

NTT FILE COPY

1

SECURITY CLASSIFICATION OF THIS PAGE

AD-A199 057

PAGE

1a. REPORT SECURITY CLASSIFICATION N00014-88-J-1013		1b. RESTRICTIVE MARKINGS									
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT <div style="border: 1px solid black; padding: 5px; text-align: center;">DISTRIBUTION STATEMENT A Approved for public release Distribution U limited</div>									
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE		4. PERFORMING ORGANIZATION REPORT NUMBER(S)									
6a. NAME OF PERFORMING ORGANIZATION The George Washington Univ.		6b. OFFICE SYMBOL (If applicable)									
7a. NAME OF MONITORING ORGANIZATION		7b. ADDRESS (City, State and ZIP Code) Sponsored Research 2121 "I" St., N.W., Suite 601 Washington, D.C. 20052									
8a. NAME OF FUNDING/SPONSORING ORGANIZATION ONR/DOD		8b. OFFICE SYMBOL (If applicable)									
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-88-J-1013		10. SOURCE OF FUNDING NOS. <table border="1"><tr><td>PROGRAM ELEMENT NO.</td><td>PROJECT NO.</td><td>TASK NO.</td><td>WORK UNIT NO.</td></tr><tr><td></td><td></td><td></td><td></td></tr></table>		PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.				
PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.								
11. CONFERENCE ON UNCERTAINTY IN ENGINEERING DESIGN		12. PERSONAL AUTHOR(S) Nozer D. Singpurwalla									
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM 11/1/87 TO 10/31/88									
14. DATE OF REPORT (Yr., Mo., Day) May 10-11, 1988		15. PAGE COUNT By Sessions- V									
16. SUPPLEMENTARY NOTATION Proceeding of Conference on Statistics in Engineering Design - Four Sessions											
17. COSATI CODES <table border="1"><tr><th>FIELD</th><th>GROUP</th><th>SUB. GR.</th></tr><tr><td></td><td></td><td></td></tr></table>		FIELD	GROUP	SUB. GR.				18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) <i>Li</i>			
FIELD	GROUP	SUB. GR.									
19. ABSTRACT (Continue on reverse if necessary and identify by block number) The Conference focused on discussions of statistics and engineering, and their interrelationships. The Conference provided an opportunity for statisticians and engineers to interact on modern statistical methods and their application to product engineering and process design. The discussions concentrated on five major topical areas which are of prime importance today in the engineering and statistics communities. These are (a) achieving quality through design, (b) achieving quality for industrial processes, (c) statistics in design for flexible automation, (d) uncertainties in design optimization, and (e) statistical issues in integrated circuit design and fabrication.											
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS <input type="checkbox"/>		21. ABSTRACT SECURITY CLASSIFICATION									
22a. NAME OF RESPONSIBLE INDIVIDUAL Helen Spencer Director, Sponsored Research		22b. TELEPHONE NUMBER (Include Area Code) (202) 994-6255									
22c. OFFICE SYMBOL A											

UIED-88

CONFERENCE
ON
UNCERTAINTY
IN ENGINEERING
DESIGN

May 10-11, 1988

Session I.

Achieving Quality
Through Design

DR. GARY MCDONALD

Chairman

General Motors Research Labs



Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By <i>per ltr</i>	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

SESSION I.

**ACHIEVING QUALITY
THROUGH DESIGN**

SESSION I. ACHIEVING QUALITY THROUGH DESIGN

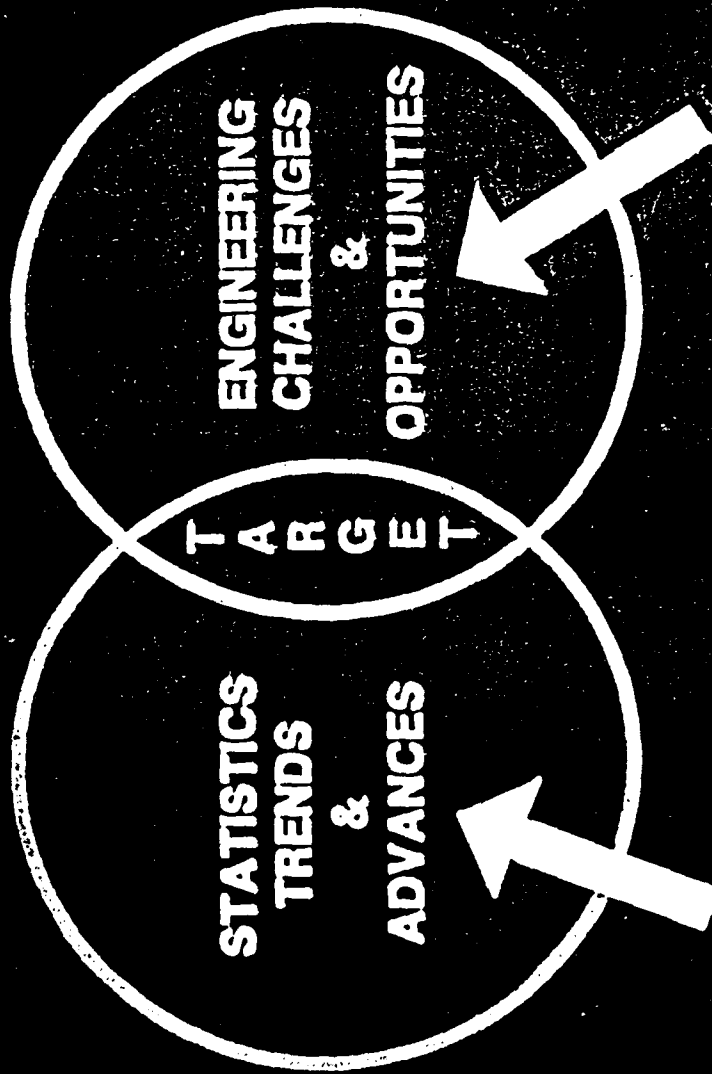
Chairman: Dr. Gary McDonald
General Motors Research Labs

Speaker: Prof. Stuart Geman
Brown University
"The Role of Statistics in
Automated Inspection and
Classification for Process Control"

Speaker: Dr. Hilarlo Oh
General Motors Corp.
"Modeling Variation to Enhance
Quality in Manufacturing"

Discussants: Stephen Fix
Sheffield Measurement
Prof. Edward J. Wegman
George Mason University

SESSION GOAL



T A R G E T

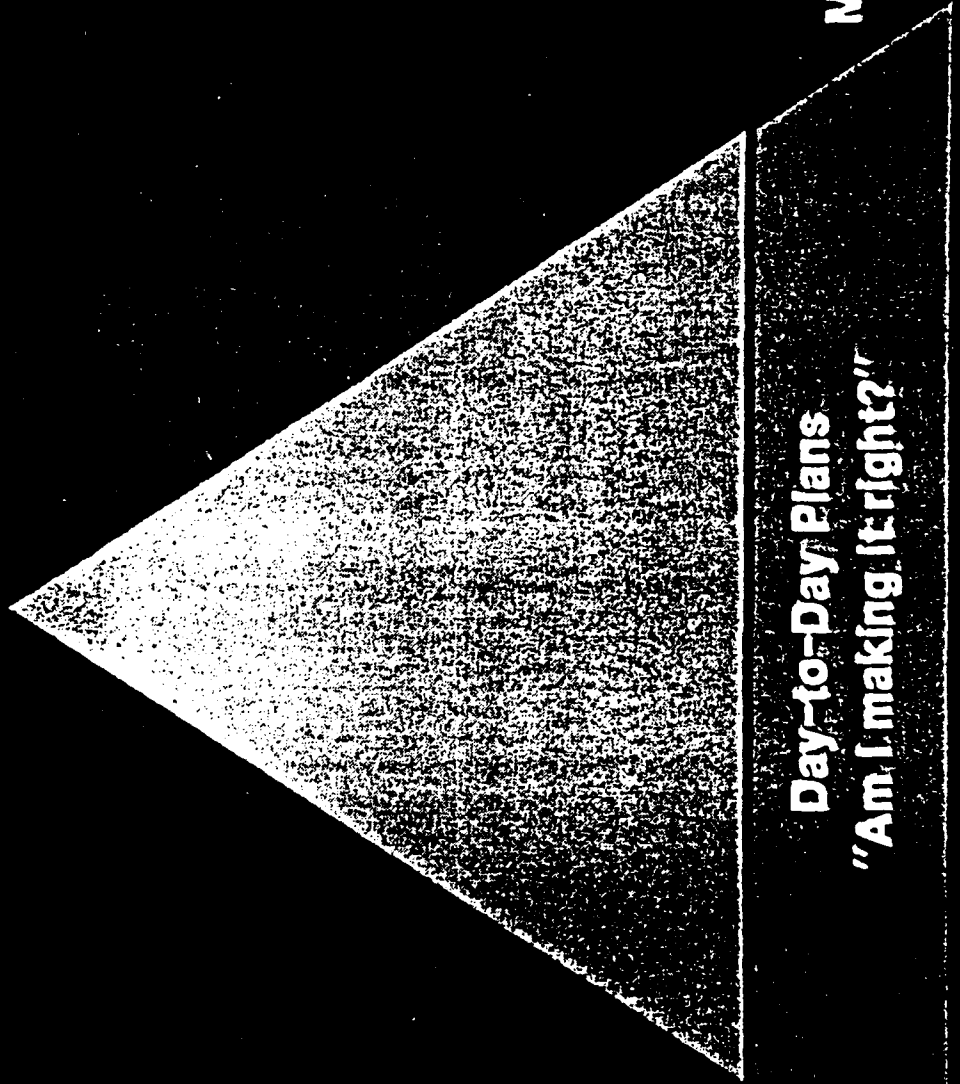
STATISTICS
TRENDS
&
ADVANCES

ENGINEERING
CHALLENGES
&
OPPORTUNITIES

- Variation Control
- Signal Processing
- Design of Experiments
- Optimization with Constraints
- • •

- Competitiveness
- Computing Advances
- Analysis vs. Experiments
- Technology Impact
- • •

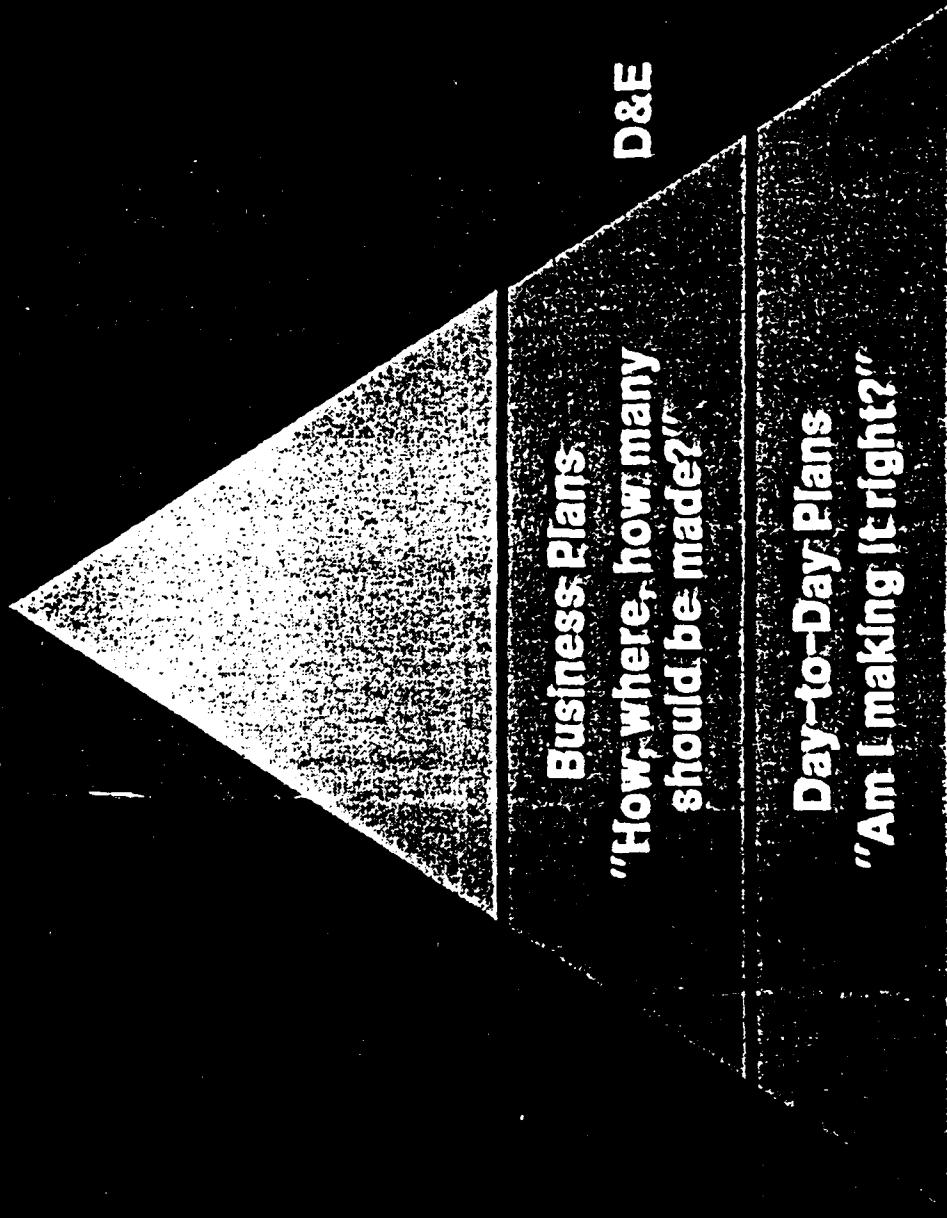
METAMORPHOSIS



Day-to-Day Plans
"Am I making it right?"

Mfg.

METAMORPHOSIS



D&E

Business Plans

"How, where, how many
should be made?"

Day-to-Day Plans

"Am I making it right?"

METAMORPHOSIS

MACRO
STATISTICS



MICRO
STATISTICS

Strategic
Plans

Planning

"What should it make?"

Business Plans

"How, where, how many
should be made?"

Day-to-Day Plans

"Am I making it right?"

THE PIMSLetter

on Business Strategy
Number 4: Product Quality

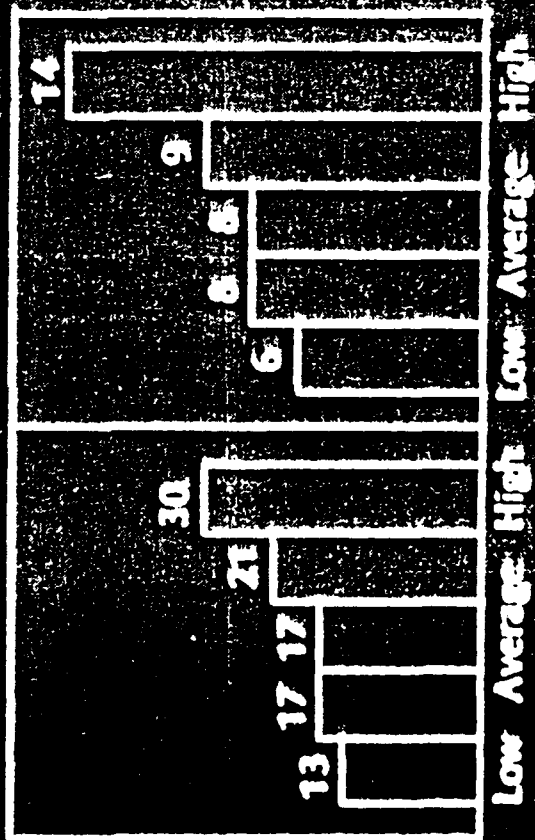
Summing Up

- High quality and high return on investment usually go together.
- Businesses offering premium-quality products usually - but not always - have large market shares and were early entrants into their markets.
- Quality is positively and significantly related to ROI for almost all kinds of products and market situations.
- A strategy of quality improvement usually leads to increased market share, but at a cost in terms of reduced short-run profitability.
- High-quality producers usually charge premium prices.

Exhibit 7

Product Quality and Profitability

Return on Investment: Net Profit, % of Sales



Product Quality

Exhibit 3

Quality and ROI by Type of Business

Type of Business	Quality Level				
	Lowest	Below Avg.	Avg.	Above Avg.	Highest
Average ROI					
Consumer Durables	16%	18%	18%	26%	32%
Consumer Non-Durables	15	21	17	23	32
Capital Goods	10	8	13	20	21
Raw Materials	13	21	21	21	35
Components	12	20	20	22	36
Supplies	16	13	19	25	36

lecture outline

THE ROLE OF STATISTICS IN
AUTOMATED INSPECTION AND CLASSIFICATION
FOR PROCESS CONTROL

Stuart Geman, Brown University

I. Introduction

- A. Systematic improvements in computing hardware have sustained hopes for "intelligent" computers.
- B. In many application areas, such as speech processing, filtering, and vision, algorithm and software development have not kept pace with hardware improvement. Partly, this is because the *appropriate* scientific tools have not been utilized.
- C. In many (most?) cases, the *mathematical sciences*, especially *probability* and *statistics* offer the right tools for constructing suitable algorithms. In this regard, the utility of "symbolic processing" and other "AI" tools such as frames and schemas have been overestimated and oversold.

D. Encouraging a role for the mathematical sciences:

1. Traditional consulting arrangements encourage shallow scientific involvement by mathematicians.
2. Probabilists, statisticians, and other mathematical scientists can best contribute by joining and initiating research efforts in high technology areas, such as speech, filtering, and vision.
3. "Neural networks" are parallel processing systems for statistical inference. This well-funded field could catalyze involvement of mathematicians, physicists, and other scientists not traditionally associated with algorithm development for "intelligent" processing.
4. Affiliate programs and entrepreneurial arrangements could encourage mathematicians to take leadership roles in technology research.

II. Computing horsepower - an illustrative example

- A. Parallel processing will play an increasingly important role in weapon systems and industrial automation.
- B. In the next few years, the most useful parallel machines will likely have modest numbers of relatively powerful processors.
- C. One example is the MTAP system, based on VHSIC technology, being developed for the Army.
- D. Other examples are the multi-DSP processor systems now coming on the market.
 - 1. One example has 4 to 12 DSP processors with extensive local memory and very high bandwidth communication.
 - 2. This system costs between \$15,000 (for 4 processors) and \$30,000 (for 12 processors).
 - 3. Each processor has a 10 (soon to be 16) megahertz clock, and performs two memory operations, one arithmetic operation, and one accumulate per clock cycle.
 - 4. Future versions will offer more processors.

-slides of architecture-

III. Opportunities for industry and mathematical sciences

A. Speech recognition

1. Most researchers acknowledge the IBM *framework* to be the most advanced.
2. Beautiful mathematical structure, accommodating vertical processing (interpretation guided segmentation), time warping, ...
3. Two severe weaknesses: top level (language) model, bottom level (signal) model.
4. Signal model: clustered feature vector from FFT or LPC (ARMA model).
 - a. Much more sophisticated and powerful tools exist in the statistics collection.
 - b. Modern theories of time series, Markov models, ... should be harnessed.
 - c. These are not off-the-shelf methods; mathematicians should be involved in implementation.

B. Filtering

1. Sample problem:

- $x_t, t = 0, 1, \dots, T$: "state" such as a calibrated setting of a stage handler .
- $y_t, t = 0, 1, \dots, T$: noisy observation of x_t
- Given $y_t, t = 0, 1, \dots, T$, estimate $x_t, t = 0, 1, \dots, T$.

2. Standard solution: Kalman filter with estimation of state parameters.

3. Existing hardware permits exploitation of more general framework.

4. Especially: should construct more realistic state models.

5. Example:

- a. Often have a priori knowledge about qualitative behavior of state process, e.g.
 - i. There exist occasional jumps.
 - ii. Between jumps, movement resembles random walk.
- b. Suggest: build Markov model of state using Gibbs representation of random fields:

$$P(x_t : 0 \leq t \leq T) =$$

$$\left(\frac{1}{Z_\lambda}\right) \exp\left\{-\lambda \sum_{t=1}^T \phi(x_t - x_{t-1})\right\}$$

- i. ϕ "engineered" to capture a priori knowledge.
- ii. λ estimated from data.
- iii. Computationally feasible!

C. Vision

1. Example: automatic defect detection and classification for wafer manufacturing process control. Wafer ID's read automatically for defect cataloguing.

-slides of wafers & wafer characters-

2. State-of-the-art
 - a. Template matching for defect detection (*sells*)
 - b. Optical character recognition: matched filter (*sells*)
 - c. Effectiveness
 - i. Both very sensitive to lighting, focus, and *normal process variations* (such as *texturing*).
 - ii. Not suitable for detailed multilayered (end-stage) inspection.

-slides of textures & texture histograms-

3. Mathematical technologies for inspection, classification and optical character recognition:
 - a. Inspection
 - i. Probabilistic analysis of texture: spatial statistics, random fields, ...

-slides of textures & texture segmentations-

- ii. Statistical estimation of normal process variations.
- iii. Combinatorial optimization of detection algorithm.
- iv. Result: full field of view inspection (512×512 pixels); submicron defect detection (2×2 pixels); 400 milliseconds.
 - processing time independent of complexity
 - *one* 8 megahertz DSP
 - Algorithm fully parallel (as are most vision algorithms)
- b. Classifier
 - i. Decision tree/recursive partition classifier (statisticians version of an expert system).
 - ii. Naturally accommodates statistical variation.
- c. Optical character recognition
 - i. Templates \rightarrow Relational Templates.
 - ii. Optimize speed by exploiting sequential decision framework for graph matching.

-slides of characters read in OCR experiments-

Technical Report, Division of Applied Mathematics & Center for Intelligent Control Systems.
To appear in *Proceedings of the 46th Session of the ISI, Bulletin of the ISI*, Vol. 52, (1987).

STATISTICAL METHODS FOR TOMOGRAPHIC IMAGE RECONSTRUCTION

Stuart Geman and Donald E. McClure
Division of Applied Mathematics
Brown University, Providence, Rhode Island, U.S.A.

1. Introduction

Interest in statistical approaches to reconstruction problems in emission computed tomography was greatly enhanced by the work of Shepp and Vardi (1982) on the use of maximum likelihood (ML) methods. There are earlier instances of suggestions to regard the reconstruction problem as a statistical estimation problem; however, the demonstration of the versatility of the approach as well as the specification of algorithms *that work* were advanced substantially by Shepp and Vardi's work.

The image reconstruction problem, viewed as an estimation problem, is inherently nonparametric: one seeks an estimate of a function of general form on a continuous domain. As such, it is widely recognized that the estimates need to be regularized or smoothed, especially in "small sample" implementations. Various approaches to regularization have been suggested, including penalized ML, the method of sieves, and Bayesian methods. In Geman and McClure (1985), we proposed that *a priori* spatial information be built into a statistical reconstruction algorithm, in a Bayesian approach, by quantifying spatial constraints in the form of a Gibbs prior distribution. In this paper we will expand on our earlier description and present recent work on parameter estimation for the Gibbs priors, which leads to completely data-driven algorithms.

This application to single photon emission computed tomography (SPECT) follows a general Bayesian paradigm for problems in image processing and vision laid out in Geman and Geman (1984) and Grenander (1984).

1. Following the general procedure, we shall describe in §2 and §3 the *deformations* that transform the object X that we wish to reconstruct into the data Y that we can observe. The deformation is embodied in a probability distribution $\Pi(Y|X)$ reflecting the physics of the observed phenomenon, the characteristics of the sensor used, etc. Alone, $\Pi(Y|X)$ is the basis for ML reconstructions.
2. The prior information about the unknown object X is then prescribed in the form of a *Gibbs prior distribution* $\Pi(X)$ (§4). In this particular application, the prior is designed to express *spatial constraints*, such as "isotope concentrations within subregions of common tissue type and common metabolic activity are fairly homogeneous."
3. The prior distribution and the deformation mechanism let us solve, by Bayes formula, for the *posterior distribution* $\Pi(X|Y)$ (§5).
4. With the posterior distribution in hand, we can base reconstruction algorithms on the statistical principle of *minimum risk*. In §5 we define procedures for the MAP and MMSE reconstructions.
5. The special association of the Gibbs prior with a statistical mechanical system translates into Monte Carlo computational methods, which mimic the dynamics of the physical system. *Stochastic relaxation* (§5) is a technique for sampling from the posterior distribution $\Pi(X|Y)$.

In §6 we describe two methods for parameter estimation for a *natural parameter* of the family of Gibbs priors. Finally, we give examples of the reconstruction and parameter estimation methods.

This paper is intended as an introduction, with emphasis on the *statistical* perspective. A more complete discussion of physical, computational, and mathematical issues will be provided in a following paper.

2. Single Photon Emission Tomography

Emission tomography is used to determine the distribution of a pharmaceutical in a part of the body such as the brain, liver, or heart. Depending upon the pharmaceutical used, this concentration can be taken as a measure of local blood flow (perfusion) and/or local metabolic activity. Glucose, for example, is taken up by neuronal cells in proportion to metabolic activity, and the latter generally mirrors recent electrical activity. Thus, areas of the brain most used in performing a cognitive or motor task will demonstrate a relatively increased uptake of glucose immediately following the task. For the heart, pharmaceuticals can be chosen whose uptake reflects local perfusion. The concentration of these pharmaceuticals can thereby be used to assess the adequacy of blood flow to the different parts of the heart.

In SPECT, pharmaceutical concentration is estimated by detecting *photon emissions* from an injected or inhaled dose of the pharmaceutical that has been chemically combined with a radioactive isotope. This combined agent is called a radiopharmaceutical. The goal of SPECT is to determine radiopharmaceutical concentration (equivalently, isotope concentration or density) as a function of position in a region of the body. Detectors with collimators are strategically placed around the region of interest, and these are able to count photons emitted by radioactive decay of the isotope. A detector will capture those photons which escape attenuation and whose trajectories carry them down the bore of the collimator.

The determination from photon counts of isotope concentration as a function of position is referred to as *reconstruction*.

Let $X(s)$ denote the concentration of the radiopharmaceutical at the point $s = (x, y)$ in the domain Ω of interest. We shall take Ω to be a bounded two-dimensional region, though for the models and methods we will describe there are no essential changes when Ω is three-dimensional.

We assume that the detectors are arranged in a linear array, at equally spaced lateral sampling intervals, and that the detector array can be positioned at any orientation θ relative to the x -axis. (See Figure 1.) We assume the detectors are of so-called *parallel bore* type, meaning that they detect only those photons in a small interval $[\theta - \Delta\theta/2, \theta + \Delta\theta/2]$ when the array has orientation θ . Let L denote the total number of detectors in the array and let $\Delta\sigma$ denote the spacing between detectors.

The physical effects incorporated in the model are the *spatial Poisson process* that describes the sites of the radioactive decays from which photons emanate and the *process of photon attenuation* by which photons are annihilated and their energy is absorbed by matter through which their trajectories pass. Attenuation is accurately described by a linear attenuation function $\mu(s)$ on Ω . The function μ is assumed to be known; values of μ for bone, muscle, etc. and for various photon energies are known *a priori* or could be measured by transmission tomographic methods. Attenuation is a memoryless process and we can thus deduce the functional form of the probability that a photon survives to reach the detector array. When a photon trajectory has direction θ and it emanates from site $s = (x, y)$ in Ω , then

$$P(\text{photon survival}) = \exp\left\{-\int_{L(x,y)} \mu(\xi, \eta) d\ell\right\},$$

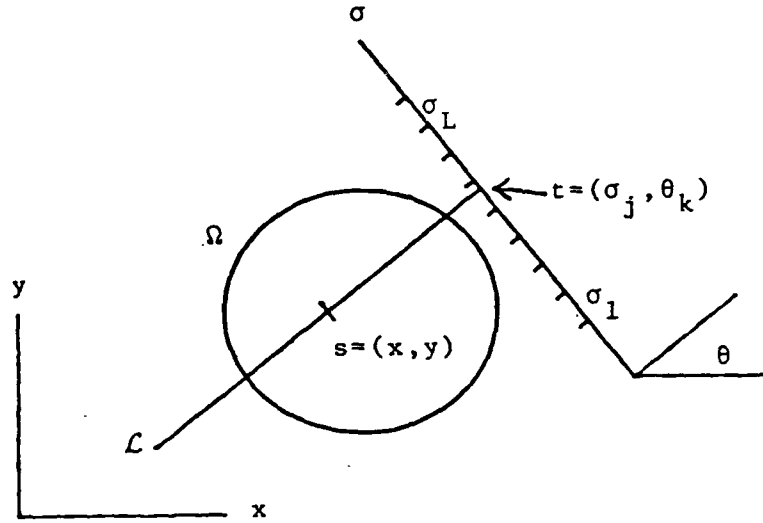


Figure 1.

where the line integral is taken over the segment $\mathcal{L}(x, y)$ from (x, y) to the detector and dl is differential arc length.

For our sampling design, we shall position the detector array at n equally spaced angles θ_k for duration T time units at each angle. Then at each angle, we observe the random variables $Y(t)$, for $t \in D_k = \{(\sigma_j, \theta_k), j = 1, \dots, L\}$ that give the numbers of photons reaching the respective detectors during the sampling interval. Assuming that (i) photons are generated by a spatially nonhomogeneous Poisson process with intensity $X(s)$ per time unit, and (ii) the orientations θ of photon trajectories are uniformly distributed on $[0, 2\pi)$, we can show that $Y(t)$, for $t \in D = \bigcup_{k=1}^n D_k$, is itself a Poisson process with a nonhomogeneous intensity function described in terms of the attenuated Radon transform (ART) of X . The ART of X is defined as

$$(R_{\mu, T}X)(\sigma, \theta) = \int_{\mathcal{L}} TX(x, y) \exp\left(-\int_{\mathcal{L}(x, y)} \mu(\xi, \eta) dl'\right) dl$$

where \mathcal{L} is the line with orientation θ , through point σ of the detector array, $\mathcal{L}(x, y)$ is the segment of \mathcal{L} starting at point (x, y) in Ω , and dl and dl' are differential arc length in the two line integrals. The intensity function of Y is then given by

$$EY(t) = \int_{\theta_k - \Delta\theta/2}^{\theta_k + \Delta\theta/2} \int_{\sigma_j - \Delta\sigma/2}^{\sigma_j + \Delta\sigma/2} (R_{\mu, T}X)(\sigma, \theta) d\sigma d\theta,$$

where $t = (\sigma_j, \theta_k)$. The important feature of this representation is that the intensity function of Y is the result of applying a positive linear integral operator \mathcal{A}_T to X :

$$EY = \mathcal{A}_T X. \quad (2.1)$$

The model includes the predominant physical effects. Other potentially significant effects, such as photon scattering and background radiation, are assumed for now to be negligible. Further, we have not included effects from the sensor, such as imperfect collimation, blurring, and noise. We note, however, that the reconstruction methods described below, since they are based on the generally applicable principles of maximum likelihood and Bayes optimality, are adaptable to models incorporating additional physical and sensor effects. Mertus (1987) has made extensions for scattering and collimation errors.

3. Maximum Likelihood and EM

A variety of reconstruction algorithms for emission tomography are described by Budinger et. al. (1979). The algorithms that are traditionally used are based on ideas of extracting a signal in the presence of noise and related methods of linear filtering.

More recently, interest has been heightened in the use of algorithms that use fuller information of the mathematical model sketched above, along with the ML principle. Shepp and Vardi (1982) laid the mathematical foundations and developed effective algorithms based on EM (Dempster, Laird and Rubin (1977)) for implementing ML reconstructions in positron emission tomography (PET). A penetrating description, written from a statistician's perspective, is given in Vardi, Shepp and Kaufman (1985). (In PET, photon attenuation does not enter the model relating isotope concentration to the observables.) McClure and Accomando (1984) have developed the foundations for applying ML to SPECT reconstructions and have implemented EM algorithms on a variety of computer systems. Independently, Miller, Snyder and Miller (1985) have made similar extensions of ML and EM for SPECT.

By exploiting properties of the Poisson process, it can be shown that the observables $Y(t)$ are mutually independent and Poisson distributed; the likelihood function is then easily obtained from (2.1). To carry out a ML reconstruction, we first discretize the domain Ω into *pixels* parameterized by discrete points s in a square lattice S . Now $\{X(s)\}_{s \in S}$ represents a piecewise constant approximation of the isotope concentration on the continuous domain. When Ω is discretized, then equation (2.1) takes the form

$$EY = A_T X,$$

where A_T is a matrix, $A_T = \{A(t, s)\}_{t \in D, s \in S}$; commonly, the order of A_T is extremely large and it may not have full column rank. Now for a given X , the Poisson probability function of Y is

$$\Pi(Y|X) = \prod_{t \in D} \frac{[(A_T X)(t)]^{Y(t)}}{Y(t)!} \exp\{-(A_T X)(t)\} \quad (3.1)$$

where our notation is making convenient abuse of the distinction between a random variable and its value.

The log-likelihood function is

$$\ln L(x) = \sum_{t \in D} \{-\ln(Y(t)!) + Y(t) \ln[(A_T X)(t)] - (A_T X)(t)\}. \quad (3.2)$$

The necessary conditions for maximizing $\ln L(X)$ obtained by setting derivatives to zero do not yield explicit solutions for a maximizing X . Nonetheless, $-\ln L(X)$ is globally convex, and the ML optimization problem conveniently adapts to the EM method. In general, $-\ln L(X)$ is not strictly convex; this is an identifiability issue related to the column rank of A_T . Conditions for strict convexity are discussed by Accomando (1984).

The EM algorithm becomes an explicit iterative reconstruction procedure. We initialize the iteration with $\{X^{(0)}(s)\}_{s \in S}$ and update $X^{(\tau)}$ by the formula

$$X^{(\tau+1)} = \{[A_T'(Y \oslash A_T X^{(\tau)})] \oslash A_T' 1\} \otimes X^{(\tau)}, \quad (3.3)$$

where 1 is the vector whose components are identically one, \oslash denotes component-by-component division, and \otimes denotes component-by-component multiplication. At each step, the iteration requires two (large) matrix multiplications. The sequence of iterates converges to an X^* that maximizes $\ln L(X)$. Consistency results that depend on the sampling design and on the discretization of Ω can be proved.

Figure 2C in §5 shows an example of a ML reconstruction for a simulation experiment. The true isotope density used for the simulated data is depicted in Panel A of Figure 2. The noisy appearance of the ML reconstruction is not atypical, even though the sample size is rather large in this experiment for estimating the 32×32 discrete image. The high degree of local irregularity occurs because ML builds in *no* spatial information, e.g. about relative locations of pixels in the grid. Snyder and Miller (1985), recognizing the inherent nonparametric nature of the reconstruction problem, have suggested using Grenander's method of sieves (Grenander (1981)) to regularize the ML estimates. Accomando (1984) also uses sieves to study consistency questions.

4. Gibbs Prior Distribution

We suggest a Bayesian formulation for incorporating prior spatial constraints into the reconstructions. We shall construct a prior distribution on X that captures simple prior expectations about the qualitative nature of the isotope density. Mainly, we wish to exploit the anticipated smoothness of X . Neighboring locations will typically have similar intensity levels. But we must also accommodate sharp changes in concentration, which might occur across an arterial wall or across a boundary between two tissue types.

In the spirit of nonparametric estimation, we might construct the prior on a suitable space of functions $X : \Omega \rightarrow \mathbb{R}$. It is more convenient, however, to do the construction on the discrete domain S introduced in §3. The prior, therefore, is on the array $X = \{X(s)\}_{s \in S}$. The range of values of $X(s)$ will be confined to a compact interval, usually $[0, 255]$, and might be further restricted to only the integer values in the interval. As a further convenience, we will restrict ourselves to priors with *Gibbs representation*

$$\Pi(X) = \frac{1}{Z} \exp \{-U(X)\} \quad (4.1)$$

where Z is the normalizing constant, $Z = \int \exp\{-U(X)\} dX$, and $U : \mathbb{R}^S \rightarrow \mathbb{R}$ is known as the "energy". As it stands, the Gibbs representation is only mildly restrictive since U is arbitrary. However, we shall restrict U to involve only "nearest neighbor" interactions among the components of X .

We employ the Gibbs representation because it is easier to design an energy function with desired properties (such as localization of interactions, Markovian restrictions on conditional distributions, ...) than it is to construct a distribution Π directly. We will design U so that the expected configurations have *low energy* as they do in a real physical system. The expected configurations are those for which typical neighboring sites $s, t \in S$ have similar intensities $X(s), X(t)$. This is a *local* constraint and it is conveniently captured by a locally composed energy function U ,

$$U(X) = \sum_{[s,t]} \beta \phi(X(s) - X(t)) + \sum_{\langle s,t \rangle} \frac{\beta}{\sqrt{2}} \phi(X(s) - X(t)). \quad (4.2)$$

Here we use $\{s, t\}$ to indicate that s and t are nearest horizontal or vertical neighbors in the lattice S , and $\langle s, t \rangle$ to denote diagonal neighbors. The constant β is positive and the function $\phi(\xi)$ is even and minimized at $\xi = 0$. Thus U is minimized by configurations of constant intensity. Under the Gibbs distribution (4.1) the more likely isotope densities are those with small site-to-site variation in intensity.

This definition of ϕ and U induces a graph on S in which each pixel site s is linked to its eight nearest neighbors in the square lattice. The distribution Π then determines a Markov random field with this neighborhood structure.

To achieve the desired properties for the more likely isotope densities, the exact form of ϕ is probably not important, but its qualitative features can make a difference. We have experimented with ϕ 's that are increasing in ξ for $\xi \geq 0$. An obvious choice is $\phi(\xi) = \xi^2$, but then under $\Pi(X)$, large intensity gradients, as would be associated with certain natural boundaries, are exceedingly unlikely. Instead, we use functions of the form

$$\phi(\xi) = \frac{-1}{1 + (\xi/\delta)^2} \quad (4.3)$$

where δ , like β , is a constant to be fixed later.

There are two free parameters in the specification of U : δ is easily interpreted as a scale parameter on the range of values of $X(s)$ and β controls the "strength" of the interactions between a pixel and its neighbors. It is a *natural parameter* of the exponential family (4.1), and admits meaningful statistical and physical interpretations. From the physical viewpoint, β is the reciprocal of temperature for the statistical mechanical system defined by (4.1). From the statistical viewpoint, it will be seen as a "smoothing parameter" controlling the tradeoff for our reconstructions between the influence of the observables and the influence of the prior constraints.

Levitin and Herman(1987) have recently proposed the use of Gaussian priors in a Bayesian formulation. Liang and Hart (1987) also suggest the use of Gaussian priors, as well as others, deduced by max-ent arguments from prior constraints on low-order moments of X . Our earlier experiments with the *quadratic* energy function indicated that the resulting Bayesian algorithms oversmoothed real boundaries where the difference $(X(s) - X(t))$ should be allowed to be large. The finite asymptotic behavior of our ϕ -function was designed to mitigate this oversmoothing.

5. Posterior Distribution and Bayes Optimal Reconstructions

From (3.1) and (4.1) the *posterior distribution* on X is

$$\Pi(X|Y) = \frac{1}{Z(Y)} \exp\{-U(X) + \sum_{t \in D} [Y(t) \ln[(A_T X)(t)] - (A_T X)(t)]\} \quad (5.1)$$

where $Z(Y)$ is a normalizing constant that depends on Y .

We have developed algorithms for two Bayes optimal reconstructions of X —the minimum-mean-squared-error (MMSE) estimator

$$X^* = E(X|Y) \quad (5.2)$$

and the maximum-a-posteriori (MAP) estimator, which maximizes the value of $\Pi(X|Y)$ or equivalently minimizes the posterior energy

$$U(X) - \sum_{t \in D} [Y(t) \ln[(A_T X)(t)] - (A_T X)(t)]. \quad (5.3)$$

The algorithm for each of these reconstructions is built around a technique for simulating computationally the dynamics of a statistical mechanical system with energy given by (5.3). Details of the generic algorithm, a variant of the Metropolis algorithm (Metropolis et. al. (1953)) known as *stochastic relaxation* (SR), are given in Geman and Geman (1984); the idea is sketched below.

Notice in (5.3) that we have the usual equivalence between Bayesian MAP estimation and so-called penalized ML. ML maximizes

$$\sum_{t \in D} [Y(t) \ln[(A_T X)(t)] - (A_T X)(t)],$$

whereas MAP estimation includes the "penalty term" $-U(X)$, which penalizes lack of smoothness. One advantage, we believe, of the Bayesian viewpoint is that it suggests mechanisms for *estimating* the required degree of smoothness, which amounts to estimating the pivotal parameter β in the Gibbs prior. We focus on this estimation problem in the next section.

MMSE Algorithm. The computational method is iterative. We initialize $X = X^{(0)}$. In practice, we choose a "good" initialization such as the EM reconstruction, but easy theory says that convergence is independent of the initialization. We visit each site s in the pixel array, successively in any order, and replace $X(s)$ by a value sampled from the conditional distribution on $X(s)$, under (5.1) and conditioning on all $X(t), t \neq s$; this is the essence of stochastic relaxation (SR). The iterates $X^{(\tau)}$ form a Markov chain with equilibrium distribution (5.1). The ergodicity of the chain guarantees that an ergodic average of $\{X^{(\tau)}\}_{\tau=0}^{\infty}$ will converge to X^* a.s. In practice, we compute N iterates and average the final M , with choices such as $N = 25$ and $M = 5$. The selection of suitable M and N can be guided by monitoring stabilization of statistics of the successive iterates $X^{(\tau)}$.

MAP Algorithm. Computing the minimum of (5.3) is, in general, a hard problem. The method of simulated annealing can be implemented to yield a sequence $\{X^{(\tau)}\}$ converging in distribution to a MAP estimator X^* . The procedure is similar to SR. The fundamental ideas are described in Pincus (1970), Cerný (1982), and Kirkpatrick, Gellatt and Vecchi (1983). See also Geman and Geman (1984) for applications to image processing.

For the design of *feasible* algorithms, we are guided by pragmatism as well as by the theoretical underpinnings of SR and simulated annealing. First we compute the ML reconstruction by EM. Then—in the language of simulated annealing—we "run" the physical system with posterior energy (5.3) at *zero temperature*. When our state-space (the range of values for $X(s)$) is discrete, this amounts to using Besag's method of Iterated Conditional Modes (ICM), Besag (1986). When the state-space is a continuous interval and the temporal index is also continuous ($\tau \in [0, \infty)$), we implement this step by performing gradient descent on (5.3) starting at the EM reconstruction. The local minimum of (5.3) obtained by ICM or by gradient descent is our approximate MAP estimate of X .

Note that ICM and gradient descent do not guarantee convergence to a global minimum of (5.3). The rationale for making a judicious choice for the initialization is to capture a "good" local minimum for the approximate MAP reconstruction.

Figure 2, Panels D, E, and F, shows approximate MAP reconstructions of the known *phantom* depicted in Figure 2A. First the ART of the phantom X was computed, for $n = 60$ sampling angles and $L = 64$ lateral sampling steps. The nonuniform attenuation function μ depicted in 2B was used to compute the ART; it builds a very substantial attenuation effect into the model. The Poisson data Y was generated to satisfy (3.1). Figure 2C shows the approximate ML reconstruction after 54

iterations of EM. For the Bayesian reconstructions, we used the prior of (4.1)–(4.3), with $\delta = 0.7$, and with the range of $X(s)$ in $[0, 15]$. Each of the Bayesian reconstructions is computed by gradient descent starting from the EM estimate (EM-GrD). The effect of different choices for β on the degree of smoothing is apparent. We shall discuss estimation of β in the next section.

6. Parameter Estimation

The choice of β is critical. With $\beta = 0$ the estimator is undersmoothed, and in fact MAP estimation is just ML, since the prior is uniform. If β is too large, the estimator is too faithful to the prior and is oversmoothed. The parameter δ is also important, though we have found that (i) its value can usually be set based on information about the range of values $\{X(s)\}$, and (ii) reconstructions are not sensitive to moderate changes in δ . The discussion here will focus on β .

Because of the setting in which reconstruction algorithms are actually used, it is desirable to design estimation methods that work with a sample \tilde{Y} of size one from the observable process. The isotope density \tilde{X} is assumed to be drawn from a Gibbs prior with unknown β , but known δ (4.3). We shall estimate β from \tilde{Y} and use the estimate $\hat{\beta}$ in the MMSE or MAP reconstruction program. It is reasonable to do this with a single observation \tilde{Y} , since \tilde{Y} contains a large amount of data about \tilde{X} , which, in turn, contains a large amount of data about the local energy function $U(X)$.

To be more explicit about the dependency on β of the prior and posterior distributions, we introduce the function

$$V(X) = \sum_{[s,t]} \phi(X(s) - X(t)) + \frac{1}{\sqrt{2}} \sum_{\langle s,t \rangle} \phi(X(s) - X(t)).$$

V is just U/β . The prior is now written

$$\Pi(X) = \frac{1}{Z_\beta} \exp \{-\beta V(X)\}$$

and the posterior, given \tilde{Y} , is

$$\Pi(X|\tilde{Y}) = \frac{1}{Z_\beta(\tilde{Y})} \exp \left\{ -\beta V(X) + \sum_{t \in D} [\tilde{Y}(t) \ln[(A_T X)(t)] - (A_T X)(t)] \right\}$$

Now $V(X)$ is a *complete-data sufficient statistic* for β . If we were able to observe \tilde{X} directly, then we could, in principle, solve the likelihood equation

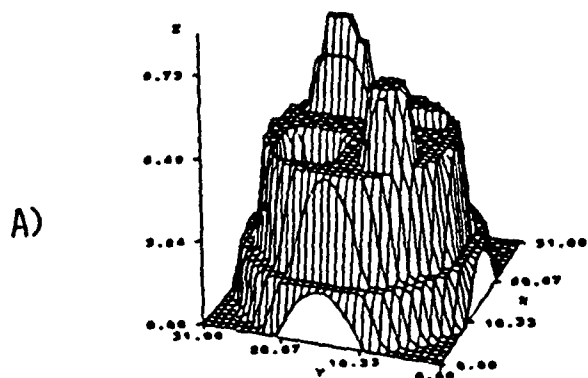
$$E_\beta[V(X)] = V(\tilde{X}) \quad (6.1)$$

for the ML estimate of β . The left-hand side of (6.1) is strictly decreasing in β and thus (6.1) yields a unique root $\hat{\beta}$.

Our situation is more complicated than this since we do not observe \tilde{X} , but instead we see only the *incomplete data* \tilde{Y} . We have a classic setup for application of EM. The EM algorithm, when it converges, will yield a root of the incomplete-data likelihood equation

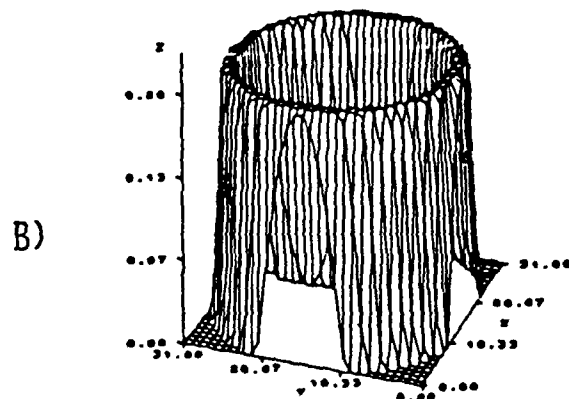
$$E_\beta[V(X)] = E_\beta[V(X)|\tilde{Y}]; \quad (6.2)$$

ISOTOPE DENSITY

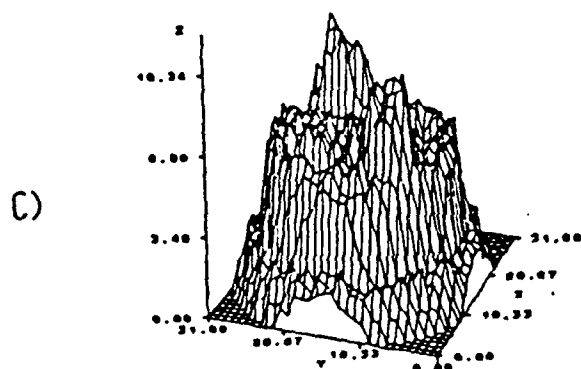


ATTENUATION FUNCTION

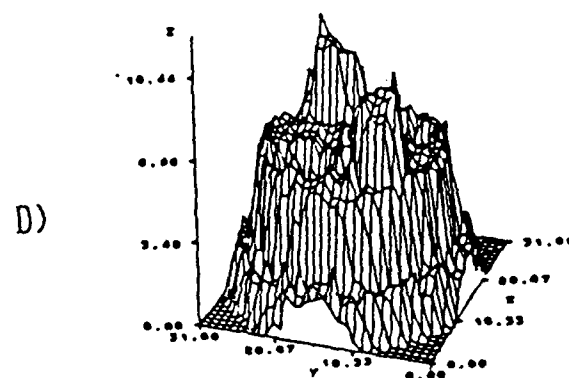
μ



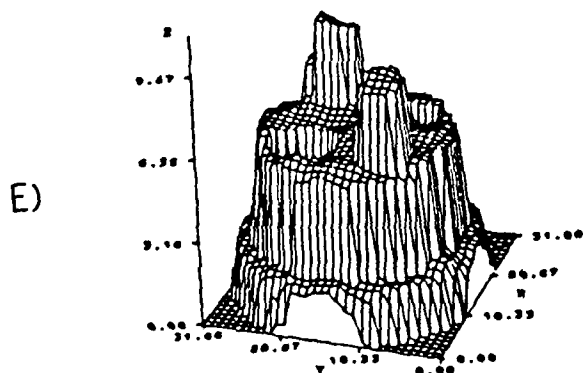
EM RECONSTRUCTION
54 ITERATIONS
 $n = 60, L = 64, \text{COUNTS} = 1,998,189$



MAP (EM-G_rD) RECONSTRUCTION
 $\beta = 0.25$



MAP (EM-G_rD) RECONSTRUCTION
 $\beta = 2.0$



MAP (EM-G_rD) RECONSTRUCTION
 $\beta = 6.0$

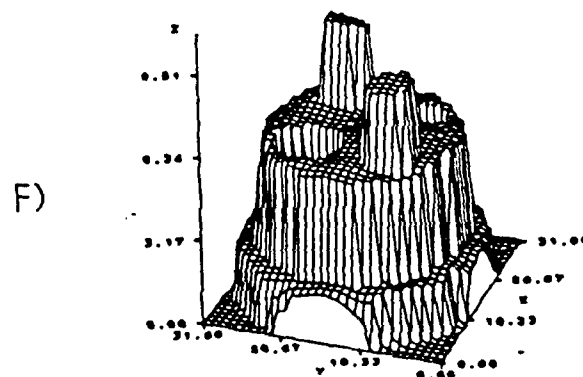


Figure 2.

see Dempster, Laird and Rubin (1977). We note that there is no proof of uniqueness of roots of (6.2). Conceptually, (6.2) is solved at the intersection of two monotone decreasing functions of β . Whether (6.2) does admit multiple solutions is an open and elusive theoretical question.

To solve (6.2), the EM algorithm consists of two alternating steps—estimation of the right-hand side of (6.2) for prescribed β (E-step) and computation of the root $\hat{\beta}$ of (6.2), substituting the current estimate of $E_{\beta}(V(X)|\tilde{Y})$ on the right-hand side. Specifically, we fix an initial $\beta = \beta^0$ and an initial $X = X^0$ (and hence V^0). Then solve

E-step. Estimate the complete-data sufficient statistic:

$$V^{(\tau+1)} = E_{\beta^{(\tau)}}(V(X)|\tilde{Y}) \quad (6.3a)$$

M-step. Determine $\beta^{(\tau+1)}$ as the solution of

$$E_{\beta}[V(X)] = V^{(\tau+1)}. \quad (6.3b)$$

The first step is done using SR, using say ten steps of SR and averaging the last five values of $V(X^{(\nu)})$. The second step is a simple root-finding calculation once the curve $E_{\beta}[V(X)]$ is known. Conveniently, the SR procedure simultaneously yields updates $X^{(\tau)}$ of the MMSE reconstruction. Thus (6.3a) and (6.3b) together give a completely data-driven method of reconstruction.

The construction of $E_{\beta}[V(X)]$ as a function of β can be done "off line", once and for all. We have done this using SR to simulate 230 configurations X from the prior (4.1) for β -values ranging from 0 to 6. Five replications were done at each of forty-six values of β . The resulting curve, fit by a cubic-spline regression function, is depicted in Figure 3. The calculation of this curve required forty-one hours of CPU time, using a highly optimized program on the 100 Megaflop Star Technologies ST100 Array Processor.

J. Mertus (1987) has developed an efficient vectorized FORTRAN program for the EM estimation/reconstruction procedure described above. Each E-step, with ten sweeps of SR, takes on the order of seven minutes of CPU time on an IBM3090 or about three minutes on a CYBER 205, working on a 64×64 pixel lattice S , for isotope densities X having their support on a disk of diameter 44 pixels (about 22cm) and with a range of 64 grey levels. (These values correspond to our real data sets.) The computational requirements are enormous, but not prohibitive.

To circumvent the computational demands of EM, we have devised and experimented with a *moment method* for estimating β . The goal is to have a direct estimation method for β that can be applied to the observable \tilde{Y} without requiring intermediate reconstruction of X . We construct a statistic $M(Y)$ based on the notion that the smoothness of Y will reflect the magnitude of β in the same way that the smoothness of X does. The exact form of $M(Y)$ is also guided by our knowledge of the Poisson distribution of Y and ability to compute theoretical moments of the Poisson random variables.

For the detector bin at angle θ_k and at sampling step σ_j , denote $t = (\sigma_j, \theta_k)$ and $t^+ = (\sigma_{j+1}, \theta_k)$. Also, introduce the notation $a(t) = (A_T \mathbf{1})(t)$, where $\mathbf{1}$ is the vector with components identically equal to one; $a(t)$ is simply the row-sum of A_T associated with the detector at location t . Then define the moment statistic

$$M(Y) = \sum_{k=1}^n \sum_{j=22}^{42} \left[\left(\frac{Y(t)}{a(t)} - \frac{Y(t^+)}{a(t^+)} \right)^2 - \frac{Y(t)}{a^2(t)} - \frac{Y(t^+)}{a^2(t^+)} \right] \quad (6.4)$$

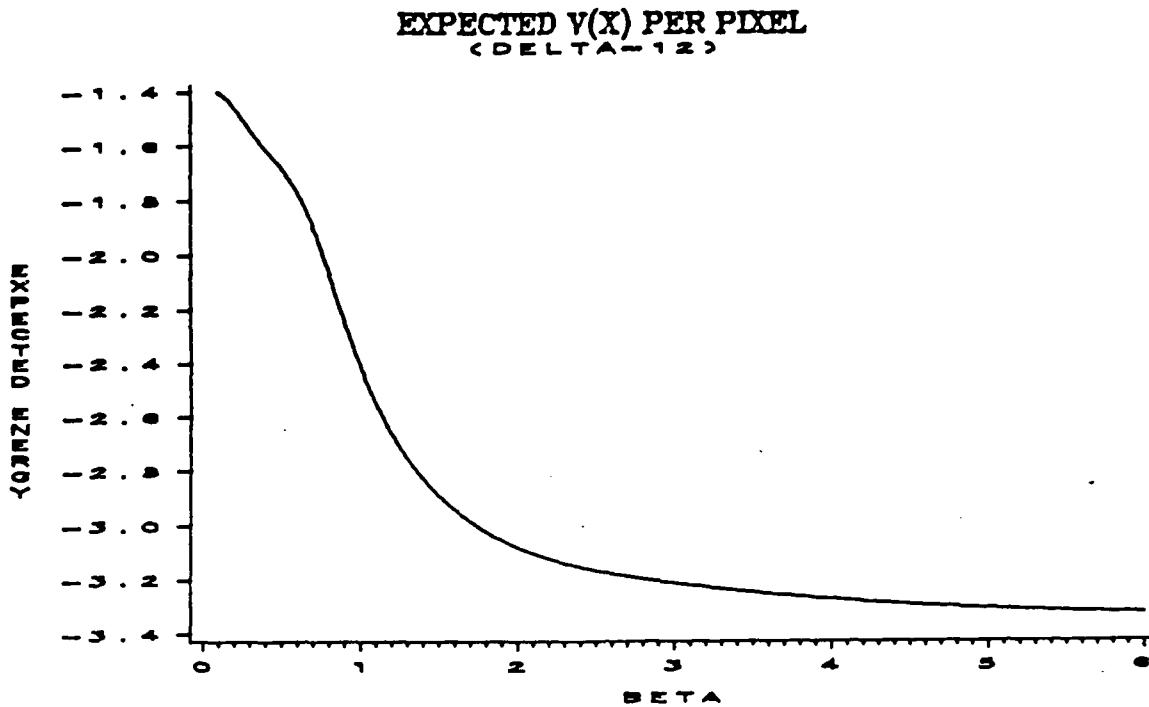


Figure 3.

The inner sum restricts the moment to the central part of the support of the isotope density to avoid edge effects. The expectation of $M(Y)$, for given X , is a measure of roughness of normalized ART projections of X :

$$E(M(Y)|X) = \sum \sum T^2 \left[\frac{A_T X(t)}{a(t)} - \frac{A_T X(t^+)}{a(t^+)} \right]^2. \quad (6.5)$$

We anticipate that the expectation $E_\beta[M(Y)]$ with respect to the prior will have the same general behavior as $E_\beta[V(X)]$ in (6.1). Accordingly, we define the moment estimate β^* of β as the root of the equation

$$E_\beta[M(Y)] = M(\tilde{Y}). \quad (6.6)$$

The effort to compute β^* is trivial, once the left-hand side of (6.6) is known as a function of β .

We have constructed the curve describing $E_\beta[M(Y)]$ using the same simulated X -data that generated $E_\beta[V(X)]$ in Figure 3. Figure 4 shows the resulting curve; it does, indeed, exhibit the same qualitative behavior as the curve in Figure 3.

A variety of experiments have been done with both the EM and moment method of estimating β . The most ideal circumstance, of course, is when the model truly fits the data.

In one such experiment, an X -array was generated from the prior (4.1) with $\beta = 1$. (As above, we used a 64×64 pixel lattice, 64 grey levels, a disk of diameter 44 pixels for the support of X , and a uniform attenuation function for the construction of A_T .) In implementing the E-step,

EXPECTED $M(Y)$ OVER CENTRAL 21 PROJECTION BINS (DELTA=12)

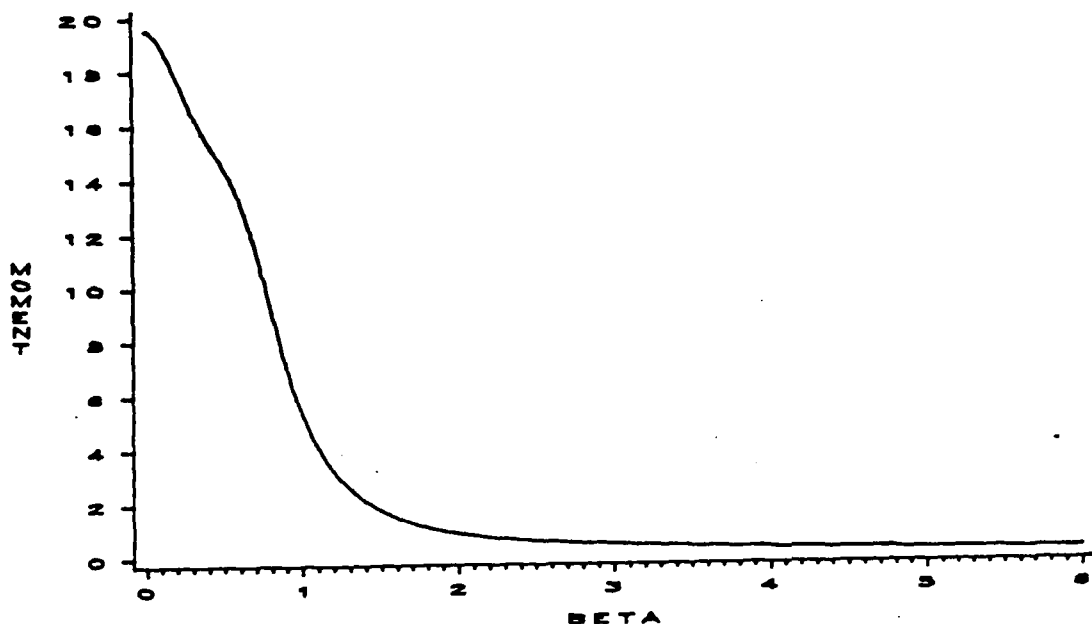


Figure 4.

ten passes of SR were performed and the last five values of $V(X^{(\tau)})$ were averaged to estimate the right-hand side of (6.3a). When $\beta^0 = 0.0$, the successive iterates of $\beta^{(\tau)}$ from the M-step were 0.63, 0.85, 0.95, 1.01, and 1.03. When $\beta^0 = 6.0$, the successive were 2.26, 1.15, 1.08, 1.05, 1.05, and 1.04. For the same X -array, five independent replications of the observable Y process were generated and the moment method yielded estimates β^* of 0.97, 1.00, 1.02, 1.06, and 0.98; the five estimates have mean 1.005 and standard deviation 0.034.

For more thorough testing of the moment method, a test set of X -arrays—independent of the set used to construct the curves in Figures 3 and 4—was generated with β -values ranging from 0.5 to 2.5. For each β , five X -arrays were generated, and for each X -array, five independent replications of the Y process were simulated. Figure 5 depicts the estimate errors $\beta^* - \beta$ for each of the twenty-five experiments at each β -value. The dispersion of the errors as a function of β is what one would anticipate from the slope of $E_\beta[M(Y)]$.

7. Reconstruction Experiments

We report on two experiments which have been run on real and simulated data to learn about the performance of the Bayesian reconstruction methods in cases for which the underlying model does not fit exactly. One simulation experiment was designed to test the versatility and robustness of the methods to known departures from the model. The other experiment illustrates the performance of the algorithms on real data from a lung section.

The pseudo-grey-level images in Figures 6 and 7 associate high values in $[0, 63]$ with black and low values with white. Our ability to present pictorial examples is limited by the printing process for this volume. Interested readers can obtain higher resolution copies of photographs on request to

ESTIMATE ERROR (MOMENT METHOD)

$$\beta^* - \beta$$

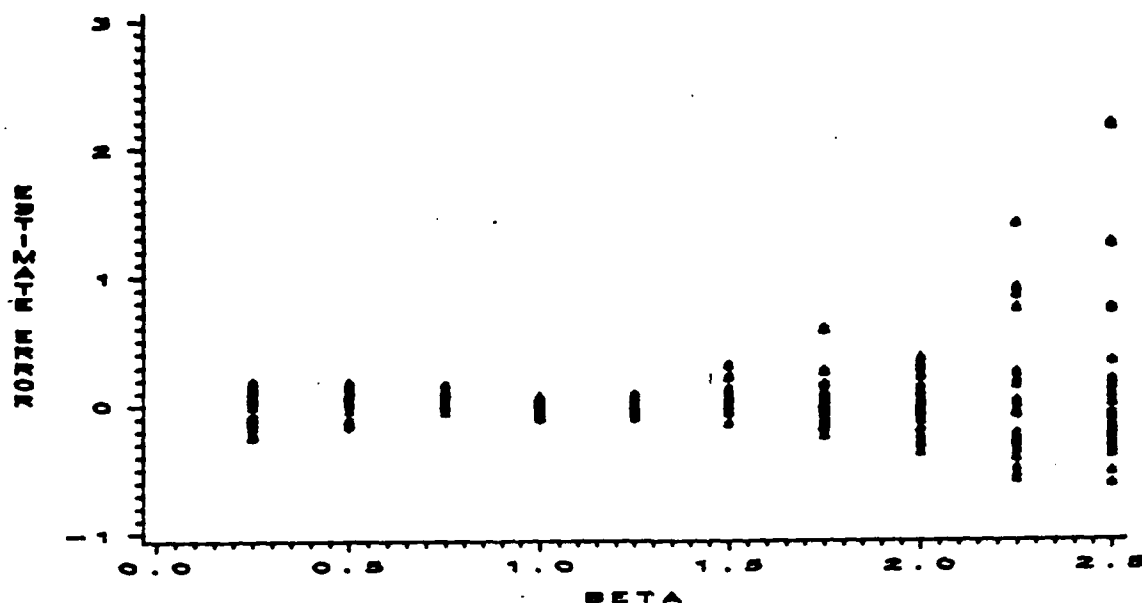


Figure 5.

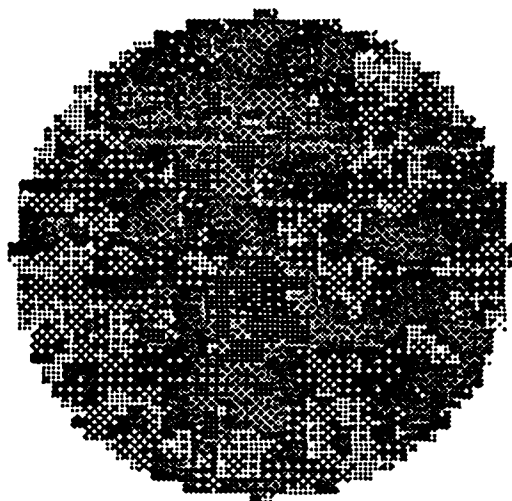
D.E. McClure.

Experiment 1. A phantom isotope density (Figure 6A) was designed to have a combination of (i) large-scale structure, including subregions of Ω with considerable differences in intensity, and (ii) local irregularity of the same qualitative nature as that of sample functions from the Gibbs model (4.1)–(4.3), yet not precisely fitting the Gibbs model. Two functions were averaged to form the phantom. First, an array with a sharp spike in intensity (near the center, below the middle) was constructed. Second, an array was sampled from (4.1)–(4.3) with parameter values $\beta = 1$, and $\delta = 12$. Intuitively, the local structure of the average will be governed by the array sampled from the Gibbs model. But observe that the rescaling of this array due to the arithmetic averaging means that it will not exactly fit a model from the same family. Roughly speaking, the averaging has the effect of smoothing the array so that it will be better described by a Gibbs model with larger β -value, assuming δ is fixed for now. We thus anticipate estimated values of β larger than the value $\beta = 1$ used to generate the Gibbsian part of the averaged phantom.

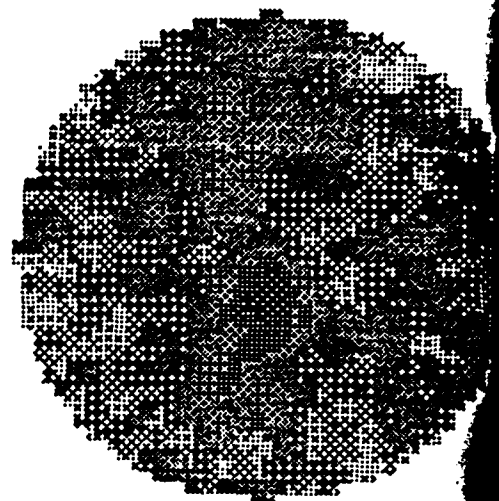
To simulate the emitted photons, the constant linear attenuation function $\mu \equiv 0.2$ was chosen, corresponding to approximately ten percent attenuation per centimeter for our scaling of the real system. A total of 663,144 photons were counted at 64 angles θ , with $L = 64$ bins on the linear detector array; in actuality, only 44 of the bins collect positive counts because the support of the phantom is contained in a smaller disk of diameter 44 pixels.

Reconstructions are depicted in Panels B–F of Figure 6. All were constructed on the range $[0, 63]$ with parameter $\delta = 12$. The MMSE reconstruction, with β estimated by EM (6.3) is shown in Panel 6B. When β was initialized at $\beta^0 = 0.0$, the successive EM iterates from (6.3b) were 0.52,

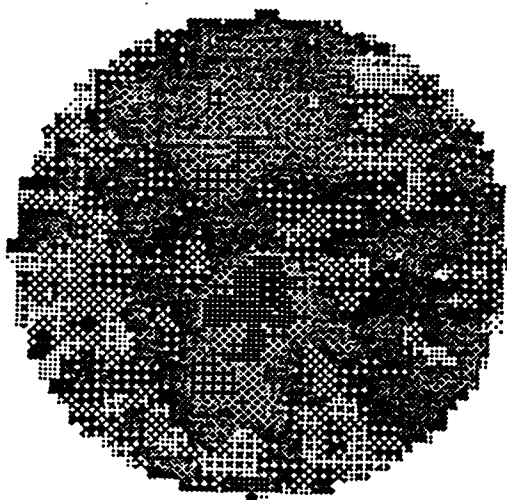
A)



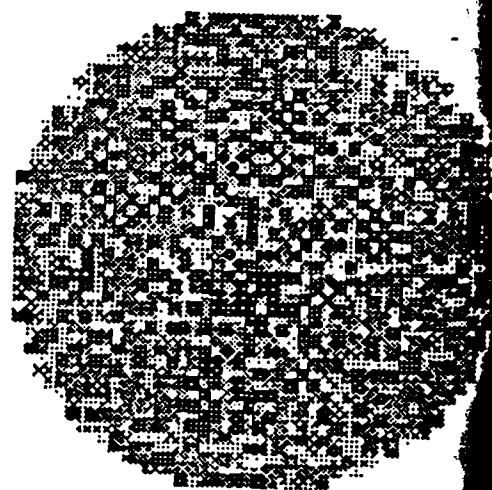
B)



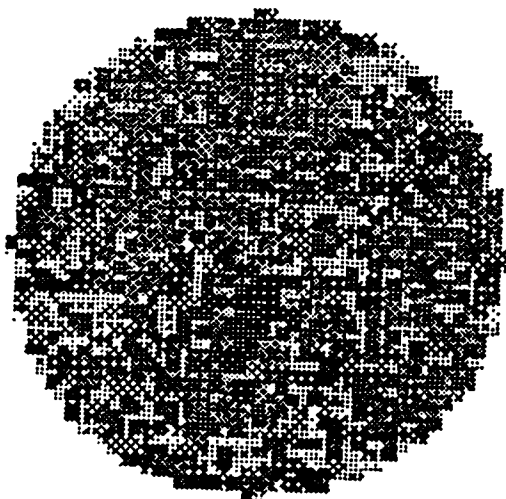
C)



D)



E)



F)

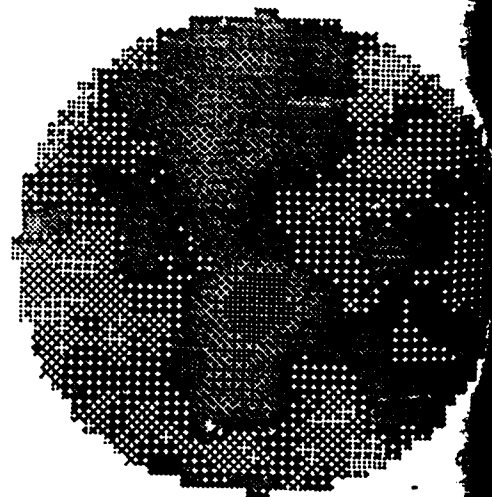


Figure 6. Simulated Data, 663,144 Total Counts.

0.72, 0.85, 1.01, 1.18, 1.29, 1.34, 1.38, 1.40, 1.41, ..., 1.47 after thirteen steps of (6.3b). The MMSE in Panel 6B was run at $\hat{\beta} = 1.47$. (The moment estimate of β was $\beta^* = 1.38$.) Panel 6C depicts an approximate MAP reconstruction obtained by ICM, with $\hat{\beta} = 1.47$ and using the MMSE in Panel 6B to initialize the local minimization of the posterior energy. Characteristically, the MAP is slightly smoother than the MMSE; on a video monitor the difference is perceptible and manifests itself in apparent coarser transitions between grey levels in the MAP image.

The EM reconstruction after 5000 (!) steps of (3.3) is shown in Panel 6D. When (3.3) is run with double precision, the successive iterates still continue to increase the log-likelihood (3.2) after 5000 iterations. Panel 6E shows an MMSE run with a value of $\beta = 0.52$, which is too small (undersmoothing). Panel 6F shows an MMSE run with a value of $\beta = 4.40$, which is too large (oversmoothing).

Experiment 2. A total of 124,136 photons were counted from a cross-section of a patient's torso, including the lungs. The observed data are depicted in the so-called sinogram in Figure 7A. The darkness in the figure is proportional to the number of detected photons. The first column of Panel 7A corresponds to the linear detector being positioned to the right of the lung section; the subinterval of high counts in this column is the "shadow" of the region of high isotope concentration in the lung. The successive columns in Panel 7A correspond, in turn, to the data from the successive sampling angles. We are using the same sampling design as in Experiment 1, with 64 equally spaced angles θ and $L = 64$ lateral sampling steps on the linear detector array.

For the reconstructions,, we set the linear attenuation function again at $\mu \equiv 0.2$. The reconstructions were done on the range $[0, 63]$ with fixed $\delta = 12$.

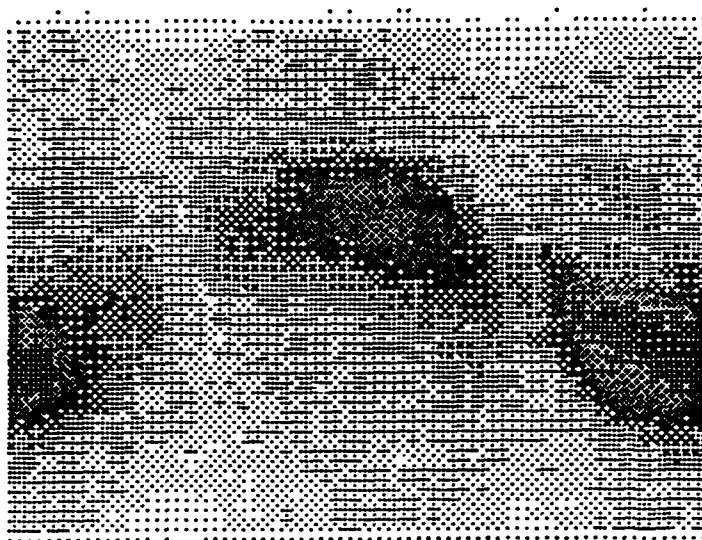
Panel 7B shows the EM reconstruction after 5000 steps of (3.3). The "hot spot" in the lung is apparent, but local structure is difficult to distinguish. Panel 7C shows the MMSE reconstruction with β estimated at $\hat{\beta} = 4.56$ after four steps of the EM estimation procedure (6.3); here we initialized $\beta^0 = 6.0$. Panel 7D shows an approximate MAP reconstruction formed by applying ICM, setting $\hat{\beta} = 4.56$, and using the EM reconstruction in Panel 7B to initialize the local minimization of the posterior energy. Again in this experiment, the MAP reconstruction is somewhat smoother than the MMSE.

The moment estimate for β in this example is $\beta^* = 2.71$. The moment estimate is sensitive to sharp singularities in the isotope concentration, such as the hot spot in the lung data. We feel that the moment method can be made more robust by using terms other than the quadratic variation used in (6.4) for the summands that define the moment statistic. There are analytical obstacles, however, to calculating a bias correction for alternative summands, so that the expectation of the moment statistic, given X , is a function of differences alone, as (6.5) is.

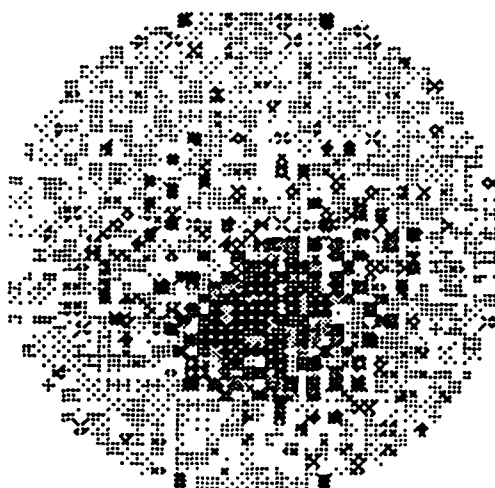
Acknowledgements

This research was supported in part by the U.S. Army Research Office Contract DAAL03-86-K-0171 to the Center for Intelligent Control Systems (Brown-Harvard-MIT), the National Science Foundation, Grant DMS-8352087, North American Philips, and the General Motors Corporation. We have also benefited from the support of the National Science Foundation Office of Advanced Scientific Computing Supercomputer Centers Program through computing resources made available at the John von Neumann Center and at the University of Illinois. John Mertus did a significant part of the experiments reported in §6 and §7. George Lorient has provided essential support in the development of algorithms and programs for the Star Technologies ST100 Array Processor.

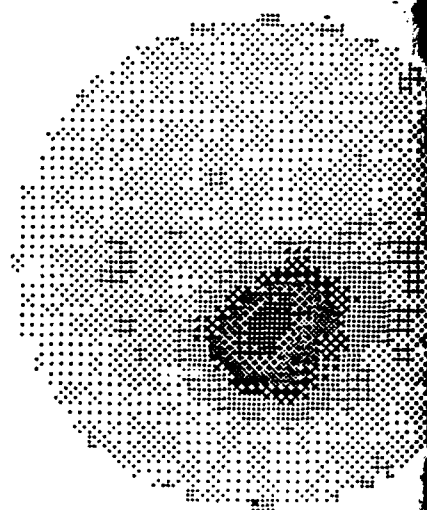
A)



B)



C)



D)

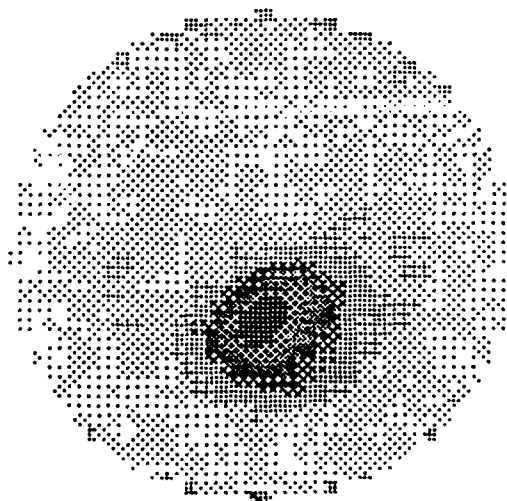


Figure 7. Real Lung Data, 124,136 Total Counts.

BIBLIOGRAPHY

1. N. Accomando (1984), "Maximum likelihood reconstruction of a two-dimensional Poisson intensity function from attenuated projections," Ph.D. Thesis, Division of Applied Mathematics, Brown University.
2. J. Besag (1986), "On the statistical analysis of dirty pictures," *J. R. Statist. Soc., B*, Vol. 48, 259-302.
3. T. Budinger, G. Gullberg, and R. Huesman (1979), "Emission computed tomography," in *Image Reconstruction from Projections: Implementation and Application*, G. Herman (editor), Springer-Verlag.
4. V. Cerný (1982), "A thermodynamic approach to the travelling salesman problem: An efficient simulation algorithm," Inst. Phys. & Biophys., Comenius University, Bratislava, preprint.
5. A. Dempster, N. Laird, and D. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., B*, Vol. 39, 1-38.
6. S. Geman and D. Geman (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. and Machine Intelligence*, 6, 721-741.
7. S. Geman and D. E. McClure (1985), "Bayesian image analysis: An application to single photon emission tomography," *Proc. American Statistical Association, Statistical Computing Section*, 12-18.
8. U. Grenander (1981), *Abstract Inference*, John Wiley & Sons, New York.
9. U. Grenander (1984), "Tutorial in Pattern Theory," Lecture Notes Volume, Division of Applied Mathematics, Brown University.
10. S. Kirkpatrick, C.D. Gellatt, and M.P. Vecchi (1983), "Optimization by simulated annealing," *Science*, 220, 671-680.
11. E. Levitan and G.T. Herman (1987), "A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography," *IEEE Trans. on Medical Imaging*, to appear.
12. Z. Liang and H. Hart (1987), "Bayesian image processing of data from constrained source distributions-I: Nonvalued, uncorrelated and correlated constraints," *Bull. Math. Biol.*, Vol. 49, 51-74.
13. Z. Liang and H. Hart (1987), "Bayesian image processing of data from constrained source distributions-II: Valued, uncorrelated and correlated constraints," *Bull. Math. Biol.*, Vol. 49, 75-91.
14. J. Mertus (1987), "Self-calibrating Bayesian methods for image reconstruction in emission tomography," Ph.D. Thesis, Division of Applied Mathematics, Brown University, in preparation.
15. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953), "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, Vol 21, 1087-1091.

16. M.I. Miller, D.L. Snyder, and T.R. Miller (1985), "Maximum-likelihood reconstruction for single-photon emission computed tomography," *IEEE Trans. on Nuclear Science*, NS-32, 769-778.
17. M. Pincus (1970), "A Monte Carlo method for the approximate solution of certain types of constrained optimization problems," *Operations Research*, 18, 1225-1228.
18. L.A. Shepp and Y. Vardi (1982), "Maximum likelihood reconstruction in positron emission tomography," *IEEE Trans. on Medical Imaging*, 1, 113-122.
19. D.L. Snyder and M.I. Miller (1985), "The use of sieves to stabilize images produced with the EM algorithm for emission tomography," *IEEE Trans. on Nuclear Science*, NS-32, 3864-3872.
20. Y. Vardi, L.A. Shepp, and L. Kaufman (1985), "A statistical model for positron emission tomography," *JASA*, Vol. 80, 8-20 and 34-37.

SUMMARY

The reconstruction problem for SPECT (single photon emission computed tomography) is formulated as a statistical estimation problem: *estimate the nonhomogeneous intensity function of a two- (or three-) dimensional Poisson process from indirect observations*. Previously, this has been addressed using the principle of maximum likelihood, but the likelihood method does not incorporate spatial constraints. Alternatively, spatial information about the unknown intensity function can be described by a Gibbs prior distribution and this then leads to Bayesian methods for the reconstruction (estimation) problem. Bayesian reconstructions are described and illustrated by examples using both real and simulated data. A parameter estimation problem for the Gibbs prior distributions is posed. Two methods are suggested and illustrated for the subsidiary parameter estimation problem. Computational algorithms are given.

RÉSUMÉ

Nous considérons le problème de reconstruction de SPECT (single photon emission computed tomography) comme étant un problème d'estimation; c'est à dire que nous estimons la fonction d'intensité (nonhomogène) d'un processus Poissonien à 2 (ou 3) dimensions. Jusqu'à maintenant, ce problème a été traité en utilisant le principe du maximum de vraisemblance; mais cette méthode ne tient pas compte des contraintes spatiales. D'autre part, l'information spatiale sur la fonction d'intensité inconnue peut être traduite par l'emploi d'une distribution de Gibbs a priori, et nous sommes conduit à une méthode Bayésienne pour le problème de reconstruction. Nous décrivons des reconstructions Bayésiennes et donnons des exemples utilisant à la fois des données réelles et simulées. Nous posons des questions sur l'estimation des paramètres de la distribution a priori de Gibbs, et nous suggérons et donnons des exemples d'application de deux méthodes pour ce problème subsidiaire de l'estimation de paramètres. Nous donnons aussi les algorithmes utilisés.

BACKGROUND ARTICLE

BOUNDARY DETECTION BY CONSTRAINED OPTIMIZATION

Donald Geman ¹

Department of Mathematics and Statistics
University of Massachusetts
Amherst, Massachusetts 01003/USA

Stuart Geman ²

Division of Applied Mathematics
Brown University
Providence, RI 02912/USA

Christine Graffigne ²

Division of Applied Mathematics
Brown University
Providence, RI 02912/USA

Ping Dong ¹

Department of Mathematics and Statistics
University of Massachusetts
Amherst, Massachusetts 01003/USA

¹ Research partially supported by Office of Naval Research contract N00014-86-K-0027.

² Research partially supported by Army Research Office contract DAAL03-86-K-0171 to the Center for Intelligent Control Systems, National Science Foundation grant number DMS-8352087, and the General Motors Research Laboratories.

ABSTRACT

We exploit a Bayesian framework, using Gibbs priors, for finding boundaries and for partitioning scenes into homogeneous regions. In both applications, the prior model is a joint probability distribution for the array of pixel grey levels and an array of "labels." In boundary finding, the labels are binary, zero or one, representing the absence or presence of boundary elements. In partitioning, the label values are *generic*: two labels are the same when the corresponding scene locations are considered to belong to the same region. The prior incorporates a measure of *disparity* between certain spatial features of pairs of blocks of pixel grey levels, using the Kolmogorov-Smirnov nonparametric measure of difference between the distributions of these features. Large disparities encourage intervening boundaries and distinct partition labels. The number of model parameters is minimized by *forbidding* label configurations that are inconsistent with prior beliefs, such as those defining very small regions, or redundant or blindly ending boundary placements. Forbidden configurations are assigned prior probability zero. We examine the MAP (*maximum a posteriori*) estimator of boundary placements and partitionings. The forbidden states introduce constraints into the calculation of MAP configurations. Stochastic relaxation methods are extended to accommodate *constrained* optimization, and experiments are performed on some texture collages and some natural scenes.

INDEX TERMS

Boundary finding, segmentation, texture discrimination, Bayesian inference, Gibbs distribution, Markov random field, MAP estimate, stochastic relaxation, annealing, constrained optimization.

1. INTRODUCTION

Most problems in image analysis, from signal restoration to object recognition, involve inference about physical entities, usually those in a three-dimensional scene. These inferences (or estimates) are based on both the observed data, usually radiant energy (or range) measurements, and general information, and imply representations of the image in terms of *unobserved* attributes or label variables. The labels may be abstract ("edge", "class k") or concrete ("occluding edge", "grass"), measurements (depth, surface normal) or semantical ("two bolts"). They may be conclusive, or serve as intermediate data structures for further analysis, perhaps involving additional data, assorted "sketches", or stored models.

This work is about two such representations: partitions and boundaries. The partition labels do not *classify*. Instead they are *generic* and are assigned to blocks of pixels; the size of the blocks (or label resolution) depends on the resolution of the data and intended interpretations. The boundary labels are just "on" or "off", and associated with an inter-pixel sub-lattice. Two specific models are constructed in Sections 2 and 3; these are instances of a "label model", which will be outlined presently, and which in turn is an application of the Bayesian paradigm in previous work ([24],[25],[27],[33]).

Both models are applied to the problem of *texture discrimination*. The data is a grey-level image consisting of textured regions, such as a mosaic of micro-textures from the Brodatz album, a patch of rug inside plastic, or radar-imaged ice floes in water. The goal is to find the regions, either by assigning generic labels to the pixels, or by constructing a boundary map, which of course avoids the microedges within the textures. The problem is more difficult than texture identification or *classification*, in which we are presented with only one texture from a given list. Discrimination can be complicated by an absence of information about the number of textures, or about the size, shape, or number of regions. In addition, the microedges within the textures may represent sharper intensity changes than those associated with the texture boundaries.

There is no effort to "model" the textures (and hence no capacity for texture synthesis). Partitioning and boundary placements are driven by the observed spatial statistics as summarized by selected features. Still, the labeling is not unsupervised because in some cases we use "training samples" to select feature thresholds; see Sections 2 and 3. We experimented with several classes of features: the well-known ones based on co-occurrence matrices ([37]) and new ones based on "directional residuals". The latter involve third and higher order distributions, the conjecture of Julesz ([42]) notwithstanding. Finally, the model enjoys some invariance properties, with respect to changes in illumination.

There are many applications for partitioning and boundary detection. Texture is a dominant feature in remotely-sensed images, and regions cannot be distin-

guished by methods based solely on shading, such as edge detectors or clustering algorithms. Specifically, for example, one might wish to determine the concentration of ice in synthetic aperture radar images of the ocean, or analyze multispectral satellite data for land use classification. Another application is to wafer inspection: low magnification views of memory arrays appear as highly structured textures, and other geometries have a characteristic, but random, graining. Many other examples and analyses of texture can be found in [16],[20],[37],[45],[48],[60],[70],[71],[72].

Texture discrimination can be regarded as the detection of discontinuities in surface composition. We also consider the problem of locating sudden changes in depth (occluding boundaries) or shape (surface creases, etc.). The idea is to define contours which are faithful to the 3-D scene but avoid the "non-physical" edges due to noise, digitization, texture, lighting, etc. Obviously, there are discontinuities, such as shadows, which are essentially impossible to distinguish from the occluding and shape boundaries, at least without information from multiple sensors or a rich knowledge base, in which case boundary *classification* becomes possible.

The complications are well-known: digital edges tend to be very "noisy", due in part to the digitization process itself, but also to de-focusing and random effects in detecting the photons. The result is a variety of pathologies: "true" boundaries suddenly disappear, spurious ones appear haphazardly, and in general the surface transitions are highly redundant.

We formulate boundary detection as a single optimization problem, fusing the detection of edges with their pruning, linking, smoothing, and so-on. The subject of edge detection is very active, and there has been considerable progress of late in designing filters based on differential operators for "optimally" detecting various "ideal" step, crease, and other edges in noise-corrupted 1-D and 2-D signals ([11],[50],[69]). Other methods detect edges after fitting smooth surfaces to the data ([36],[34],[59]), and still others ([3],[4],[13],[51],[52],[56],[68]) perform surface reconstruction and boundary detection at the same time, and are cast in a framework similar to the set-up in [25].

The use of boundary maps as the input to further processing is ubiquitous in computer vision; for example, algorithms for stereopsis, optical flow, and simple object recognition are often based on matching boundary segments. Other applications include the analysis of medical images (e.g. angiograms and ultrasound); automated navigation [9]; and the detection of the paths of roads and geologic faults, or the edges of lakes, flood plains, and crop fields, in remotely-sensed images.

Bayesian Framework. Most would agree that a coherent theoretical framework for image analysis would support more robust and more powerful algorithms for restoration and interpretation. In this work we continue exploring an approach based on Bayesian image models, well-defined principles of inference, and a Monte Carlo computation theory. Exploiting this framework, or Bayesian paradigm, we have obtained encouraging results in several areas of application, including im-

age restoration [25] and analysis [33], computed tomography [27], and texture and boundary analysis [21],[24],[26],[30]. Other researchers have adopted and expanded this (and closely related) methodologies. For example, the application in [57] to scene segmentation based on optical flow incorporates both temporal and global interactions and a degradation model based on sensor optics and other physical principles. Additional examples include surface reconstruction [51],[52], scene segmentation based on shading and texture [15],[17], and frame-to-frame matching for computing optical flow and stereo disparity [44]. Similar models have appeared in recent work on neural networks [38], speech [7], and remote sensing [46].

The approach is Bayesian because we construct prior probability models for both observed and unobserved scene attributes. These models express the regularities and preferred relations found in most real scenes, such as the unlikeliness of "blind" endings to boundaries, or very small or thin regions, and the likeliness of meaningful transitions at discontinuities of various spatial statistics. These regularities are rarely deterministic; they are best expressed as correlations and likelihoods, and we are led to the representation of our prior expectations by a "prior" probability distribution to capture the tendencies and constraints that characterize the particular scene of interest. Inference can then be guided by this prior distribution together with a model for the *degradation*, which determines the relation between the image attributes and the observation, usually in the form of a conditional density of the latter given the former. If these steps are well-conceived, there are severe, but *appropriate*, limits imposed on the plausible restorations or interpretations.

To set the stage for the applications to partitioning and boundary finding, we shall briefly review the formal description of this Bayesian framework, as it may be applied to image processing, and make some specializations and extensions that will be needed later. This will be self-contained, but we refer to [25] for more complete discussion.

We will represent by x the (high dimensional) vector of relevant image attributes, including, for example, the digitized pixel grey levels and the zero or one (off or on) boundary labels. The *prior distribution*, Π , is a probability for x : $0 \leq \Pi(x) \leq 1 \forall x$, $\sum_x \Pi(x) = 1$, where \sum_x is summation over all configurations of x (all assignments of grey levels and boundary placements, for example). We adopt the *Gibbs representation*, which is to say that we represent Π as

$$\Pi(x) = \frac{1}{z} \exp\{-U(x)\}, \quad z = \sum_x \exp\{-U(x)\}$$

The real-valued function U is called the *energy*, and evidently determines Π . The Gibbs distribution describes the equilibrium of a physical system, suitably unconstrained, that has energy U (after an appropriate scaling) as a function of the state, x . The analogy suggests using U as a vehicle to construct Π : design an energy U that is "small" for those configurations that are compatible with prior beliefs, but

is "large" when these beliefs are violated. Then the likely states under Π will be the ones that meet with prior expectations. Building U is more natural, and can be much easier, than directly building Π (see [2],[25],[27],[33],[52],[58] for explicit examples).

The object of interest is x ; we *define* it to include the relevant attributes for the particular image processing task at hand (such as the boundary labels or the generic region labels used herein, the texture classification labels used in [30], isotope intensities for computed tomography [27], or the object classifications used in [33]). There is a problem-specific degradation that precludes directly observing x . It may be the blur and noise introduced in infrared imaging, the attenuated Radon transform that figures into emission tomography, or simply an occlusion, as when pixel grey levels are observed uncorrupted, but the object of interest is the boundary placements. In the last example, the data comprises only those components of x that correspond to pixel intensities; the actual boundary labels are of course unobserved. We will denote the *data (observations)* by y . Its components are usually pixel grey levels, but could also be, for example, range data from laser radar, or gamma camera counts from an emission tomography machine. The details of the imaging mechanism define the degradation, which we formally model by specifying the conditional distribution of y (the observation) given x (the "true" state): $\Pi(y|x)$.

Given the prior ($\Pi(x)$), the observation model ($\Pi(y|x)$), and the data (y), the *posterior distribution*, $\Pi(x|y)$, is derived by Bayes' formula:

$$\Pi(x|y) = \frac{\Pi(y|x)\Pi(x)}{\sum_{x'} \Pi(y|x')\Pi(x')}$$

It is useful to preserve the formal connection with statistical mechanics, and so we write the posterior distribution in the Gibbs representation:

$$\Pi(x|y) = \frac{1}{\hat{z}} \exp\{-\hat{U}(x)\}$$

Of course, the *posterior energy* \hat{U} , and the new normalizing constant, \hat{z} , may both depend on y , but this is fixed by observation.

The goal is to estimate x , which may correspond to restoring a blurred and noise corrupted picture, placing boundaries, classifying textures, or perhaps labeling objects, depending on the task at hand. Mostly, we have worked with two estimators, the maximum a posteriori (MAP) estimator and the posterior mean. The MAP estimator is any mode of the posterior distribution:

$$\hat{x} = \arg \max_x \Pi(x|y),$$

which is the Bayes estimator corresponding to the zero-one loss function

$$L(x, \hat{x}) = \begin{cases} 1, & \text{if } x \neq \hat{x} \\ 0, & \text{otherwise} \end{cases}$$

On the other hand, the posterior mean

$$\hat{x} = \sum_x x \Pi(x|y)$$

minimizes the mean squared error, corresponding to the loss function $L(x, \hat{x}) = |x - \hat{x}|^2$.

The appropriate loss function is necessarily problem-specific. For tomography [27] we mostly use the posterior mean, although simulations from the posterior distribution are also very informative. For the boundary and generic region labels discussed here we find MAP most appropriate.

There is skepticism about the MAP estimator: see e.g. [2],[19],[52]. It has sometimes been found to be too "global", leading to gross mislabeling in certain classification problems and "over-smoothing" in surface reconstruction and image restoration. (See [10] for a different view.) The discussion paper of Besag [2] has shed much light on the subject; see especially the remarks of Silverman [65] on MAP vs. simulations from the posterior $\Pi(x|y)$, and the remarkable comparisons between the *exact* MAP estimate and approximations derived from simulated annealing in the commentary of Greig, Porteous, and Seheult [32]. However, pixel-based error measures are too local for boundary analysis. In particular, the Bayes rule based on misclassification error rate, namely the *marginal* (individual component) modes of $\Pi(x|y)$, is unsuitable because this estimator lacks the fine structure we expect of boundary maps; placement decisions cannot be based on the data alone - pending labels (i.e. *context*) must be considered. See [52],[63], and [73] for discussions of alternative loss functions and performance criteria.

Actually computing samples, means, and modes is usually impossible, at least with today's hardware. For approximations, we use a variation of the Metropolis algorithm [53] that we call stochastic relaxation (SR). This is a highly parallel Monte Carlo algorithm that loosely simulates the approach to equilibrium of an imagined system with energy \hat{U} . Later, we will have more to say about SR and certain extensions, and a full account can be found in [23] and [25]. For now, suffice it to say that, asymptotically at least, SR can be used to sample from the posterior, or to compute its mean and mode.

We have found it convenient, especially when working with boundaries and partitionings, to extend this framework by allowing "infinite energies" (zero probabilities) in the prior distribution. (See Moussouris [55] for an analysis of Gibbs measures with "forbidden" states.) Rather than inhibiting, by high energy, "blind" boundary endings and redundant boundary representations, or a partitioning into excessively small or thin regions, we simply *disallow*, or *forbid*, these configurations. Later, we will define a function $V(x)$ that essentially counts the number of subconfigurations among the labels that are forbidden. More generally, we let $V(x)$ be

nonnegative and consider the Gibbs prior on the (allowed) set $\{x : V(x) = 0\}$:

$$\Pi(x) = \frac{1}{z} \delta_{\{V=0\}}(x) \exp\{-U(x)\}, \quad z = \sum_x \delta_{\{V=0\}}(x) \exp\{-U(x)\}$$

Whatever the degradation model, the posterior distribution will be similarly restricted, and of the form

$$\Pi(x|y) = \frac{1}{\hat{z}} \delta_{\{V=0\}}(x) \exp\{-\hat{U}(x)\}, \quad \hat{z} = \sum_x \delta_{\{V=0\}}(x) \exp\{-\hat{U}(x)\}$$

The constraint, $V(x) = 0$, amounts to a placement of infinite energy barriers in the "energy landscape". These inhibit the free flow that is essential to the good performance of SR; indeed, the theory will in general break down, and convergence is no longer guaranteed. A simple and effective solution is to introduce these barriers gradually during the relaxation process. This will be made precise in §4, with the supporting convergence theory, which is quite straightforward, layed out in [23].

We now specialize to the partitioning and boundary placement applications, in which the relevant attributes are pixel grey levels and labels, the latter either representing boundary elements or regions. To make this explicit we write $x = (x^L, x^P)$, where x^L is the vector of boundary or region labels, and x^P is the vector of pixel grey levels. Two rather different kinds of considerations will go into constructing the prior. These will be discussed in detail shortly, but the upshot is that we separate the prior energy into a pixel-label interaction term and a pure label contribution. The former, $U(x^L, x^P)$, promotes placements of boundaries, or assignments of distinct labels, between regions in the image that demonstrate distinct spatial patterns. The pure label contribution is to inhibit "blind" endings of boundaries, redundant boundary representations, small regions, and other unexpected label configurations. As discussed previously, the simplest way to avoid these unwanted configurations is to *forbid* them by introducing $V = V(x^L)$ and concentrating on $\{x : V(x^L) = 0\}$. The prior, then, is of the form

$$\Pi(x) = \frac{1}{z} \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, x^P)\}$$

As for the degradation, in this paper we shall concentrate on the common situation in which our observations of the pixel grey levels are essentially uncorrupted: $y \doteq x^P$. There is no significant blur or noise, and hence no need for grey level restoration. Our only interest is in estimating the unobserved label process x^L . $\Pi(y|x)$ is singular, and the posterior reduces to

$$\Pi(x|y) = \frac{1}{\hat{z}} \delta_{\{x^P=y\}}(x^P) \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, x^P)\},$$

$$\hat{z} = \sum_{x^P, x^L} \delta_{\{x^P=y\}}(x^P) \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, x^P)\}$$

Since $x^P = y$ is fixed by observation, we equivalently treat

$$\Pi(x^L|y) = \frac{1}{\hat{z}} \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, y)\},$$

$$\hat{z} = \sum_{x^L} \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, y)\}$$

derived from the *prior*

$$\Pi(x^L, y) = \frac{1}{z} \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, y)\},$$

$$z = \sum_{x^L, y} \delta_{\{V=0\}}(x^L) \exp\{-U(x^L, y)\}.$$

The superscript L is now superfluous, and we will henceforth simply use x when referring to the label process.

Label Model: General Form. Let $x = \{x_s, s \in S\}$ and $y = \{y_{ij}, 1 \leq i, j \leq N\}$ denote, respectively, the labels and the data; thus x_s is the label at "site" $s \in S$ and y_{ij} is the grey-level at pixel (i, j) . The set S of label sites is a regular lattice, distinct from that of the pixels, and typically more sparse; the coarseness depends on the *label resolution* σ . For partitioning, we associate each site $s \in S$ with a block of pixels, "sitting below it," if we were to stack the label lattice on top of the pixel lattice. In the boundary model, *pairs* of nearby sites in S define boundary *segments*, and these are associated with pairs of pixel blocks, sitting "across from each other," with respect to the segments (see Figure 6). Later, we will define a *neighborhood system* for S such that the bonding is nearest-neighbor (relative to σ) in the boundary model, whereas in the region model there are interactions at all scales. This has important consequences for the distribution of local minima in the "energy landscape"; see §2. Other energy functionals with global interactions can be found in [27],[31], and [57].

The "interaction" between x and y is defined in terms of an energy function

$$U(x, y) = \sum_{\langle s, t \rangle} \Psi_{s, t}(x) \Phi_{s, t}(y)$$

The summation extends over all "neighboring pairs" (or "bonds") $\langle s, t \rangle$, $s, t \in S$; $\Phi_{s, t}(y)$ is a measure of the *disparity* between the two blocks of pixel data associated with the label sites $s, t \in S$. $\Psi_{s, t}(x)$ depends only on the labels x , and x_t . In fact, we simply take $\Psi_{s, t}(x) = 1 - x_s x_t$ in the boundary model and $\Psi_{s, t}(x) = \delta_{x_s, x_t}$ in the partition model. In this way, in the "low energy states", large disparities ($\Phi > 0$)

will typically be coupled with an active boundary ($x_s = x_t = 1$) or dissimilar region labels ($x_s \neq x_t$) and small disparities ($\Phi < 0$) will be coupled with an inactive boundary ($x_s = 0$ or $x_t = 0$) or equal region labels ($x_s = x_t$).

The interaction between the labels and the data is based on various *disparity measures* for comparing two (possibly distant) blocks of image data. These measures derive from the raw data as well as from various *transformations*. We experiment with several transformations $y \rightarrow y'$ for texture analysis, for example transformations of the form

$$(1.1) \quad y'_t = |y_t - \sum \alpha_j y_{t_j}|$$

where $\sum \alpha_i = 1$ and $\{t_j\}$ are pixels nearby to pixel t and in the same row, column, or diagonal. We call these "directional residuals", regarding $\sum \alpha_j y_{t_j}$ as a "predictor" of y_t . We have also experimented with a variety of transforms suggested in [37]. Disparity is measured by the *Kolmogorov-Smirnov statistic* (or *distance*), a common tool in nonparametric statistics which has desirable invariance properties. (In particular, using the directional residuals, the disparity measure is invariant to linear distortions ($y_{ij} \rightarrow ay_{ij} + b$) of the raw data, and using the raw data itself for comparisons, the disparity measure is invariant to *all* monotone (data) transformations.) The general form of the Φ term is then

$$\Phi_{s,t}(y) = \phi\left(\max_{1 \leq i \leq m} \rho(y_{(s)}^{(i)}, y_{(t)}^{(i)})\right)$$

where ϕ is monotone increasing, ρ denotes a distance based on the Kolmogorov-Smirnov statistic (see §2), and $y_{(s)}^{(i)}, y_{(t)}^{(i)}$ are the data in the two blocks associated with $\langle s, t \rangle$ for the i^{th} transform. Often, we simply take $m = 1$ and $y^{(1)} = y$.

Apparently some of these ideas have been kicking around for a while. For example, the Kolmogorov-Smirnov statistic is recommended in [64], and reference is made to still earlier papers; more recently, see [72]. Moreover, the *distributional* properties of residuals (from surface-fitting) are advocated in [34],[61] for detecting discontinuities. It is certainly our contention that the statistical warehouse is full of useful tools for computer vision.

The other component in the model is a *penalty function* $V(x)$ which counts the number of "taboo patterns" in x ; states x for which $V(x) > 0$ are "forbidden". For example, boundary maps are penalized for dead-ends, "clutter", density, etc. whereas partitions are penalized for too many transitions or regions which are "too small".

Given the observed image y , the MAP estimate $\hat{x} = \hat{x}(y)$ is then any solution to the constrained optimization problem

$$(1.2) \quad \text{minimize}_{x: V(x)=0} U(x, y)$$

We seek to minimize the energy of the data-label interaction over all possible non-forbidden label states x .

The rationale for constrained optimization is that our expectations about certain types of labels are quite precise and rigid. For example, most "physical boundaries" are smooth, persistent, and well-localized; consequently it is reasonable to *impose* these assumptions on image boundaries, and corresponding restrictions on partition geometries. Contrast this with other inference problems, for example restoring an image degraded by blur and noise. Aside from constraints derived from scene-specific knowledge, the only reasonable *generic* constraints might be "piecewise continuity", and generally the degree of ambiguity favors more flexible constraints, such as those in U , or in the energy functions used in [25] and [27].

As mentioned earlier, the search for \hat{x} is by a version of stochastic relaxation which incorporates rigid constraints. The theoretical foundations are laid out in [23], although there is enough information provided here to keep this paper self-contained; see §4. Basically, we simulate annealing ([12], [47]) by introducing a control parameter t corresponding to "temperature", and another control parameter λ , corresponding to a Lagrange multiplier for the constraint $V = 0$. More specifically, let

$$U_k(x) = t_k^{-1} [U(x, y) + \lambda_k V(x)]$$

where y (the data) is fixed, $t_k \searrow 0$, and $\lambda_k \nearrow \infty$. The algorithm generates a sequence of states $\hat{x}_k, k = 1, 2, \dots$, by Monte Carlo sampling from the local conditional distributions of the Gibbs measures with energy functions U_k . Under suitable conditions (see §4), the sequence \hat{x}_k "converges" to a solution of (1.2).

The algorithm is computationally demanding but has the same potential for parallel implementation as standard stochastic relaxation. The experiments here were performed on serial machines but required considerably less processing time than those in [25], for example, due to lower resolution labels, departures from the "correct" annealing schedules, and deterministic approximations akin to those in [2], [18], and [27]. A "fast annealing" algorithm is reported in [67]. In any event, too much fuss over CPU times may be ill-advised. Software engineers know that it is often possible to achieve order-of-magnitude speed-ups by some modest reworkings and compromises when dedicating a general purpose algorithm to a specific task, and this has certainly been our experience. Besides, advances in hardware are systematically underestimated. It is reported in [58] that experiments in [25] requiring several hours of VAX 11/780 time were reproduced in less than one minute on the ICL DAP, and the authors speculate about real-time stochastic relaxation.

There are no multiplicative parameters in the model, such as the "smoothing" or "weighting" parameters in [3], [21], [24], [25], and [27]; in effect, the energy is $U + \lambda V$ with $\lambda = \infty$. Thresholds must be selected for the disparity measures, but much of this can be data-driven (see §5), and fortunately the performance is not

unduly sensitive to these choices within some range. Other inputs include the label resolution, block sizes, and penalty patterns. The model is robust against these choices as well, so long as modest information about the pixel resolution is available.

There are close ties with "regularization theory" for "computational vision" ([52],[62]) and even closer ones with "variational" approaches, such as those of Mumford and Shah [56], Blake [3], Blake and Zisserman [4], and Terzopoulos [68], which incorporate discontinuity constraints and penalties. Perhaps the main difference is the separation of the energy components into terms corresponding to the prior and degradation models; in particular, we regard the energy function as the (negative) log-likelihood of the posterior distribution and our optimization procedures are strongly motivated by this viewpoint. Stochastic relaxation permits us to analyze the posterior *distribution*, revealing its likely and unlikely states. For instance, the posterior mean, $E(x|y)$, is not a property of the energy per se, but is an excellent estimate in some cases ([27]). And, identifying the "regularization term" as the (negative) log-likelihood of the *prior* provides a statistical framework for estimating the regularization parameter ([27]), as well as other parameters in the model ([26],[22],[74]).

Sources of Information. All information bearing on the labeling is encoded in the posterior distribution, or, equivalently, in the (posterior) energy function and constraints. There is no pre- or post-processing. The final estimate $\hat{x} = \hat{x}(y)$ is totally a function of the model and the data. In particular, if the energy does not account for any *global* image attributes, e.g. templates or semantical variables, then there is no "top-down" or "goal-directed" component to the search process. Such is the case in this work; we are currently investigating the capacity of this methodology for integrating "high-level" information.

On the other hand, Markov random field (equivalently, Gibbs) priors have proven well-suited to *cooperative processing*. For example, several tasks can be effectively linked, such as simultaneous surface interpolation and boundary-finding [51],[52], or simultaneous filtering and deconvolution [25]. (See also [57].) More to the point, a single, complex task may involve a number of sub-procedures. For example, boundary detection involves seeding, organization, and smoothing. These sub-procedures are usually performed sequentially; here they are fully coupled.

As we have already mentioned, we consider only one data source, a single frame of visible light or L-band synthetic aperture radar. It would be desirable to incorporate data from motion, multiple views, or multiple sensors, and we are currently studying an expanded version of these models utilizing both optical and range data for boundary classification and other applications. See §6 for additional remarks about generalizations of the model and [54] for a thoughtful discussion about multivariate data.

2. PARTITION MODEL

Partitionings. Denote the pixel (image) lattice $\{(i, j) : 1 \leq i, j \leq N\}$ by S_I and let S_L (formally S in §1) be the label lattice, just a copy of S_I in the case of the partition model. For each experiment, a *resolution* σ is chosen, which determines a sub-lattice $S_L^{(\sigma)} \subset S_L$, and the coarseness of the partitioning. Larger σ 's will correspond to coarser partitionings and give more reliable results (see §5), but they lose boundary detail. Specifically, let

$$S_L^{(\sigma)} = \{(i\sigma + 1, j\sigma + 1) : 0 \leq i, j \leq \frac{N-1}{\sigma}\}.$$

Recall that the observation process, or data, consists of grey levels $y_s, s \in S_I$. With the usual grey-level discretization, the state, or configuration, space for the data is

$$\Omega_I = \{\{y_s\} : s \in S_I, 0 \leq y_s \leq 255\}$$

The configuration space for the partitioning, x , is determined by σ , and by a maximum number of allowed regions, P :

$$\Omega_L^{(\sigma, P)} = \{\{x_s\} : s \in S_L^{(\sigma)}, 0 \leq x_s \leq P-1\}.$$

Recall that the labels are generic: x defines a partitioning by identifying sites with a given label $(0, 1, \dots, P-1)$ as belonging to the same region. Only the sub-lattice $S_L^{(\sigma)}$ is labelled, and a maximum number of labels (regions) is fixed *a priori*. A prior estimate of the number of distinct (but not necessarily connected) regions must be available, since the model often subdivides homogeneous regions when P is too large (see §5). The boundary model (§3) is more robust in this regard.

Each label site $s \in S_L$ is associated with a square block $D_s \subset S_I$ of pixel sites centered at s . (Recall that S_L is just a copy of S_I ; we sometimes use " s " ambiguously to reference a site in S_L and the corresponding site in S_I .) x_s labels the pixels in $D_s : \{y_r : r \in D_s\}$. As we will see shortly, the partitioning is based on the spatial statistics of these (overlapping) sub-images. The size of D_s is therefore important. We have experimented only with textures (the boundary model has been applied more generally), and it is obvious that for these the pixel blocks $\{D_s\}_{s \in S_L^{(\sigma)}}$ must be large enough to capture the characteristic pattern of the texture, at least in comparison to the other textures present. Of course, "large enough" is with respect to the features used, but in the absence of a *multiscale* analysis, an *a priori* choice of scale is unavoidable. In all of our experiments, $|D_s| = 441$, a 21×21 square block of pixels. There is again a resolution issue: larger blocks characterize the textures more reliably, having less within-region variation, but boundary detail is sacrificed.

Label-Data Interaction. We establish a neighborhood system on the label lattice $S_L^{(\sigma)}$: each $s \in S_L^{(\sigma)}$ is associated with a set of neighbors $N_s \subset S_L^{(\sigma)}$. The

system is symmetric, meaning that $s \in N_r \Leftrightarrow r \in N_s$. As we shall see, the neighborhood system largely determines the computational burden. For now we will proceed as though the neighborhood system is given, but we will have much more to say about it shortly.

Let $\langle s, t \rangle_\sigma$ denote a neighbor pair, meaning $s, t \in S_L^{(\sigma)}$, $s \in N_t$. We will introduce a *disparity measure* $\Phi_{s,t} = \Phi_{s,t}(y)$ for each neighbor pair $\langle s, t \rangle_\sigma$. Roughly speaking, $\Phi_{s,t}$ measures the similarity between the pixel grey levels in the two pixel blocks associated with $s, t \in S_L^{(\sigma)}$. For the partition model, $\Phi_{s,t}$ is simply -1 ("similar") or +1 ("dissimilar"); it is more complicated for the boundary model. The interaction energy is then

$$U(x, y) = \sum_{\langle s, t \rangle_\sigma} \delta_{\{x_s = x_t\}} \Phi_{s,t}(y)$$

In the low energy states, similar (resp. dissimilar) pairs, $\Phi_{s,t} = -1$ (resp. $\Phi_{s,t} = +1$), are associated with identical (resp. distinct) labels: $x_s = x_t$ (resp. $x_s \neq x_t$). Although $U(x, y)$ is conceived of as the interaction term in a *prior* distribution that is jointly on x and y , only the posterior distribution is actually used, and y is fixed by observation. It would be interesting, and perhaps instructive (see [43]), to sample from the joint distribution, but computationally very expensive.

Neighborhood System. A simple example will serve to highlight the issues. Suppose y has R constant grey-level (untextured) regions ($y_s \in \{0, 1, \dots, R-1\}$, $s \in S_I$), and $\sigma = 1$ (full resolution). Of course, in this case y is a labelling, so there is no point in bringing in the partition process x ; but this is just an illustrative example. The obvious disparity measure is simply $\Phi_{s,t} = -1$ if $y_s = y_t$, and +1 otherwise:

$$(2.1) \quad U(x, y) = \sum_{\langle s, t \rangle_1} \delta_{\{x_s = x_t\}} (\delta_{\{y_s \neq y_t\}} - \delta_{\{y_s = y_t\}}).$$

Entertain, for the time being, a nearest neighbor system on S_L , which is the natural choice. To be concrete, take N_s to be the four (two horizontal and two vertical) nearest neighbors of s . There are three essential difficulties with this choice of neighborhood system. Two can be readily appreciated:

- (See Figure 1.) If $R = 2$ and $P = 3$, and if region "0" (i.e. $\{s \in S_I : y_s = 0\}$) is split into two disjoint pieces by region "1" (i.e. $\{s \in S_I : y_s = 1\}$), then (2.1) has two kinds of global minima: correct labellings, in which there are two populations of labels corresponding to the two grey-level regions; and spurious labellings, in which the three regions (two of type "0" and one of type "1") are given three distinct labels.
- (See Figure 2.) If $R = 3$, and region "0" does not neighbor region "2", then there are again two kinds of global minima: correct labellings have three labels; spurious labellings have only two, incorrectly identifying regions "0" and "2".

0	0	0	1	1	0	0	1	1	1	2	2	1	1	2	2	2	1	1	0	0
0	0	0	1	1	0	0	1	1	1	2	2	1	1	2	2	2	1	1	0	0
0	0	1	1	1	0	0	1	1	2	2	2	1	1	2	2	1	1	1	0	0
0	0	1	1	0	0	0	1	1	2	2	1	1	1	2	2	1	1	0	0	0
0	1	1	0	0	0	0	1	2	2	1	1	1	1	2	1	1	0	0	0	0
0	1	1	0	0	0	0	1	2	2	1	1	1	1	2	1	1	0	0	0	0
1	1	1	0	0	0	0	2	2	2	1	1	1	1	1	1	1	0	0	0	0

Left: original "image" and a correct labelling

Middle: a correct labelling

Right: spurious labelling

FIGURE 1

0	0	0	1	1	2	2	1	1	1	2	2	0	0	0	0	0	1	1	0	0	
0	0	0	1	1	2	2	1	1	1	2	2	0	0	0	0	0	1	1	0	0	
0	0	1	1	1	2	2	1	1	2	2	2	0	0	0	0	0	1	1	1	0	0
0	0	1	1	2	2	2	1	1	2	2	0	0	0	0	0	0	1	1	0	0	0
0	1	1	2	2	2	2	1	2	2	0	0	0	0	0	0	0	1	1	0	0	0
0	1	1	2	2	2	2	1	2	2	0	0	0	0	0	0	0	1	1	0	0	0
1	1	1	2	2	2	2	2	2	2	0	0	0	0	0	0	0	1	1	1	0	0

Left: original "image" and a correct labelling

Middle: a correct labelling

Right: spurious labelling

FIGURE 2

Quite obviously, the model requires more *global interactions*. In particular, just a few long range interactions would disambiguate the correct from the spurious labellings. Only a correct labelling would achieve the *global* minimum of U in these two examples.

The third difficulty with local neighborhoods is computational, and is already apparent when $R = 1$ and $P = 2$. This time there are only two global minima, and each is a desirable labelling ($\{x_s = 0 \forall s \in S_L\}$ or $\{x_s = 1 \forall s \in S_L\}$). But, with $N = 512$, for example, consider the label configuration in which $x_{ij} = 0$ whenever $1 \leq i \leq 256$ and $x_{ij} = 1$ whenever $257 \leq i \leq 512$, a half "black" and half "white" picture. This is a local minimum, and rather severe in that it would take very many "uphill" or "flat" moves (single site changes) to arrive at either of the global minima. SR is a *local* relaxation algorithm, and despite the various

convergence theorems, the practical fact of the matter is that "wide" local minima such as these are impossible to cope with. (But, there are some encouraging results in the direction of *multiscale* relaxation, see [29],[6],[66].) Many readers will be reminded of the Ising model, in the absence of an external field, and the notorious difficulty of finding its (two) global minima by Monte Carlo relaxation. In fact, the $R = 1$, $P = 2$ energy landscape is *identical* to that of the Ising model, as is readily demonstrated by a suitable transformation of the label variables. (Indeed, the same goes for the $R = 2$, $P = 2$ case, although this is less obvious. A suitable transformation identifies the two Ising minima with the two acceptable labellings: $x_s = 0 \Leftrightarrow y_s = 0$ and $x_s = 1 \Leftrightarrow y_s = 0$.)

These local minima can be mostly eliminated by introducing long range interactions in the label lattice $S_L^{(\sigma)}$; the same remedy as for the label ambiguities. We will provide a heuristic argument for the important role of long range interactions in creating a favorable energy landscape. In any case, simulations firmly establish their utility. First recall that the *distance* between two sites in a graph is the smallest number of edges that must be crossed in travelling from one site to the other. Notice that in the four nearest neighbor graph (two dimensional lattice) the average distance between sites is large. Correct partitioning requires *all* pairs of label sites to resolve their relationships ("same" or "different"), as dictated by the statistics of their associated pixel blocks. Of course most pairs are not neighbors. With a *local* relaxation, such as SR, the resolution is achieved by propagating relationships through intervening sites. Thus the task is facilitated by minimizing the number of intervening sites, and a relatively small number of long range connections can drastically reduce the typical number of these.

The largest distance over all pairs of sites is the *diameter* of a graph. In an appropriate limiting (large graph) sense, *random graphs* have minimum diameter among all graphs of fixed degree ¹. In light of our heuristics, this suggests a random neighborhood system for $S_L^{(\sigma)}$. Indeed, random neighborhoods have a remarkable effect on the structure of local minima for these systems. In a series of experiments, with "perfect" disparity data (such as the $\sigma = 1$ grey-level problems discussed above) we could always achieve the global minimum by single-site iterative improvement when adopting a random graph neighborhood configuration, using rather modest degrees for large graphs. *We conjecture, but have been unable to prove, that even with the degree a vanishingly small fraction of the graph size, random graphs (in the "large graph limit") have no local minima, under the Ising potential or the potential $U(x,y)$ with perfect disparity data (2.1).*

Of course the disparity data is not usually perfect. In challenging texture discrimination tasks there will be pixel blocks from the same texture that are measured as dissimilar ($\Phi_{s,t} = 1$) and others from distinct textures that are measured as

¹ A graph has *fixed degree* if each site has the same number of neighbors. The degree is then the number of neighbors per site.

similar ($\Phi_{s,t} = -1$). Under these circumstances it helps to also have near neighbor interactions, since these tend to bond neighboring label sites and thereby increase the effective number of long range interactions per site. Although near neighbors were not always needed to get the best results, we settled on using four near neighbors, and sixteen random neighbors per site, in each of our experiments (see §5). By "near neighbors" we mean the closest two horizontal and two vertical neighbors whose associated pixel blocks do not overlap. For example, with $\sigma = 7$, and using 21×21 blocks, the near neighbors have two intervening sites in $S_L^{(\sigma)}$. Details on the generation of the (pseudo) random neighbors can be found in [30]. Overall, perhaps the most effective neighborhood system would have a gradual fall-off of interaction densities with distance, a system with an equal number of neighbors at each Manhattan distance, for example.

Kolmogorov-Smirnov Statistic. At the heart of the partitioning and boundary algorithms is a disparity measure $\Phi_{s,t}$. Recall that if s, t are neighbors in $S_L^{(\sigma)}$ ($< s, t >_\sigma$) then $\Phi_{s,t}$ is a measure of disparity between two corresponding blocks of pixel data, $\{\{y_r\} : r \in D_s\}$ and $\{\{y_r\} : r \in D_t\}$ in the case of the partition model. We base $\Phi_{s,t}$ on the Kolmogorov-Smirnov distance, a measure of separation between two probability distributions, well-known in statistics. When applied to the sample distributions (i.e. histograms) for two sets of data, say $v^{(1)} = \{v_1^{(1)}, v_2^{(1)}, \dots, v_{n_1}^{(1)}\}$ and $v^{(2)} = \{v_1^{(2)}, v_2^{(2)}, \dots, v_{n_2}^{(2)}\}$ it provides a test statistics for the hypothesis that $v^{(1)}$ and $v^{(2)}$ are samples from the same underlying probability distribution, meaning that $F_1 = F_2$ where, for $i = 1, 2$, $v^{(i)} = \{v_1^{(i)}, v_2^{(i)}, \dots, v_{n_i}^{(i)}\}$ are independent and identically distributed with $F_i(t) = P(v_j^{(i)} \leq t)$. The test is designed for continuous distributions and has a powerful invariance property which will be discussed below.

The sample distribution function of a data set $\{v_1, v_2, \dots, v_n\}$ is

$$\hat{F}(t) = \frac{1}{n} \#\{k : v_k \leq t\}, \quad -\infty < t < +\infty$$

Thus, \hat{F} is a step function, with jumps occurring at the points $\{v_k\}$. It characterizes the histogram. Now consider two sets of data $v^{(1)}, v^{(2)}$ with sample distribution functions \hat{F}_1, \hat{F}_2 . The Kolmogorov-Smirnov distance (or statistic) is the maximum (vertical) distance between the graphs of \hat{F}_1, \hat{F}_2 , i.e.

$$(2.2) \quad d(v^{(1)}, v^{(2)}) = \max_{-\infty < t < +\infty} |\hat{F}_1(t) - \hat{F}_2(t)|$$

We write $d(v^{(1)}, v^{(2)})$ to emphasize the data (which in our case consists of blocks of possibly transformed pixel intensity values); the conventional notation is $d(\hat{F}_1, \hat{F}_2)$.

The invariance property is the following. Suppose $v^{(1)}, v^{(2)}$ are samples from *continuous* distributions F_1, F_2 . Then under the ("homogeneity") hypothesis $F_1 = F_2$, the probability distribution of d (as a random variable) is independent of the

(common) underlying distribution. Basically, this stems from the fact that d is invariant to strictly monotone transformations of the data, i.e.

$$d(v^{(1)}, v^{(2)}) = d(q^{(1)}, q^{(2)})$$

where $q_j^{(i)} = \tau(v_j^{(i)})$ and τ is strictly increasing or decreasing. Thus, in two sample tests for homogeneity, one rejects the null hypothesis that $F_1 = F_2$ if $d(v^{(1)}, v^{(2)}) \geq d^*$, where d^* depends only on n_1 and n_2 , and on the significance level of the test.

For our purposes, the data $v^{(1)}$ and $v^{(2)}$ consist of either the (raw) grey levels, or (in most cases of texture discrimination) transformations of these, restricted to blocks of pixels; these blocks are adjacent in the boundary model, but may be well separated in the partition model (recall that we employ a largely random topology). In either case, the assumptions made in statistical testing are generally violated: it may be unreasonable to assume that the grey levels in a block of pixels represent independent and identically distributed observations from some underlying probability distribution (although this is occasionally done). Of course, the size of the blocks relative to the image structures is very important. The blocks may contain hundreds of pixels, but if they are still small relative to the image structures, then the formal assumption will be more nearly satisfied. At any rate, the formal theory is primarily motivational. The distance (2.2) is an effective "measure of homogeneity" which is invariant to pointwise (monotone) data transformations induced by lighting and other factors.

Disparity Measures. Sometimes, just grey level histograms are enough for good partitionings, as with the SAR image of water and ice (see §5). In these cases, disparity is measured as follows. Recall that D_s , $s \in S_L^{(\sigma)}$, is a square block of pixel sites (always 21×21 in the partitioning experiments) centered at s . Let $y(D_s) = \{y_r : r \in D_s\}$. Given (possibly distant) neighbors $s, t \in S_L^{(\sigma)}$, we define $\Phi_{s,t}$ using the Kolmogorov-Smirnov statistic and a threshold c :

$$\Phi_{s,t} = 2\delta_{\{d(y(D_s), y(D_t)) > c\}}(y) - 1$$

In other words, $\Phi_{s,t}$ is 1 or -1 depending on whether the Kolmogorov-Smirnov statistic is above threshold or not.

Of course, many distinct textures have nearly identical histograms (see [26] for some experiments with partitioning and classification of such textures, also in the Bayesian framework). In these cases, discrimination will rely on features, or transformations, that go beyond raw grey levels, involving various *spatial* statistics. We use several of these at once, defining $\Phi_{s,t}$ to be 1 if the Kolmogorov-Smirnov statistic associated with *any* of these transformations exceeds a transformation-specific threshold, and -1 otherwise. The philosophy is simple: If enough transformations are employed, then two distinct textures will differ significantly in at least one of the aspects represented by the transformations. Unfortunately, the implementation

of this idea is complicated; more transformations mean more thresholds to adjust, and more possibilities for "false alarms" based on "normal" variations within homogeneous regions.

Proceeding more formally, let Λ denote one such data transformation and put $y' = \Lambda(y)$, the transformed image. In general, y'_t is a function of both y_t and the grey-levels in a window centered at $t \in S_I$. For example, y'_t might be the mean, range, or variance of y in a neighborhood of t , or a measure of the local "energy" or "entropy". Or, y'_t might be a *directional residual* defined in (1.1); *isotropic residuals*, in which the pixels $\{t_j\}$ surround t , are also effective. Notice that any Λ given by (1.1) is linear in the sense that if $y_s \rightarrow ay_s + b$, $\forall s \in S_I$ then $y'_s \rightarrow |a|y'_s$, $\forall s \in S_I$, and recall that the Kolmogorov-Smirnov statistic is *invariant* with respect to such changes. This invariance is shared by other features, such as the mean, variance, and range. It should also be noted that these transforms are decidedly *multivariate*, depending (statistically) on the marginal distributions of the data of at least dimension three. Many approaches to texture analysis are based solely on the one- or two-dimensional marginals, i.e. the grey-level histogram and co-occurrence matrices. We were not able to reliably detect some of the boundaries between the Brodatz microtextures with these standard features. Perhaps the jury is still out.

Given a family of transformations, $\Lambda_1, \Lambda_2, \dots, \Lambda_m$, we define

$$(2.3) \quad \Phi_{s,t} = \max_{1 \leq i \leq m} \left[2\delta_{\{d(y^{(i)}(D_s), y^{(i)}(D_t)) > c_i\}} (y^{(i)} - 1) \right]$$

where $y^{(i)} = \Lambda_i(y)$, $1 \leq i \leq m$, and $y^{(i)}(D_r) = \{y_s^{(i)}, s \in D_r\}$, $r = s, t$. The thresholds c_1, \dots, c_m are chosen to limit the percentage of "false alarms" (cases of exceeding threshold for pairs of blocks within the same texture); see §5.

The disparity measure (2.3) inherits the aforementioned invariance to linear shifts for many transforms, including all "differences of averages". More importantly, perhaps, imagine we are comparing two pairs of image blocks, each pair in a different region of the image. Then, roughly speaking, the two distances are automatically calibrated, regardless of the differing statistical properties of the two regions; i.e. the disparity measure has the same interpretation anywhere in the image.

Penalties. Recall that $V(x)$ counts the total number of "penalties" associated with $x \in \Omega_L^{(\sigma, P)}$. There are two kinds of "forbidden" configurations that give rise to penalties: roughly, these correspond to very *small regions* and very *narrow regions*. Fix σ and $s \in S_L^{(\sigma)}$. Let E_s be the 5×5 block of sites in $S_L^{(\sigma)}$ centered at s . A configuration x is "small at $s \in S_L^{(\sigma)}$ " if fewer than nine labels in $\{x_t : t \in E_s\}$ agree with x_s . Notice that a right corner at s is allowed; there are exactly nine

agreements. The total number of penalties for "small regions" is

$$(2.4) \quad \sum_{s \in S_L^{(\sigma)}} \delta_{\{(\sum_{t \in E_s} \delta_{\{x_t = x_s\}}) < 9\}}$$

Obviously, the numbers "5" and "9", as well as other penalty parameters below, are quite arbitrary, and could reasonably be scale-dependent.

As for "thin regions", these are regions that have a horizontal or vertical "neck" that is only one label-site wide (at resolution σ). Let τ_h be a one-site horizontal translation within $S_L^{(\sigma)}$, and let τ_v be the analogous vertical translation. Penalties arise when either $\{x_{s-\tau_h} \neq x_s, \text{ and } x_s \neq x_{s+\tau_h}\}$ or $\{x_{s-\tau_v} \neq x_s, \text{ and } x_s \neq x_{s+\tau_v}\}$. The number of "thin-region" penalties is therefore

$$(2.5) \quad \sum_{s \in S_L^{(\sigma)}} \delta_{\{x_{s-\tau_h} \neq x_s, x_s \neq x_{s+\tau_h}\}} + \delta_{\{x_{s-\tau_v} \neq x_s, x_s \neq x_{s+\tau_v}\}},$$

and $V(x)$ is just the sum of (2.4) and (2.5).

Summary. We are given

- (i) a grey-level image $y = \{y_{ij}\}$;
- (ii) a resolution $\sigma = 1, 2, \dots$, and a maximum number of labels, P ;
- (iii) a disparity measure $\Phi_{s,t}(y)$ for each pair $\langle s, t \rangle_\sigma$ in the sub-lattice $S_L^{(\sigma)}$;
- (iv) a collection of penalty patterns.

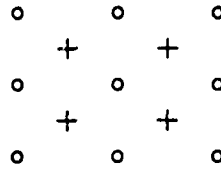
The (MAP) partitioning $\hat{x} = \hat{x}(y)$ is then any solution $x \in \Omega_L^{(\sigma, P)}$ of the constrained optimization

$$\text{minimize}_{x: V(x)=0} \sum_{\langle s, t \rangle_\sigma} \delta_{\{x_s = x_t\}} \Phi_{s,t}(y)$$

where $V(x)$ is the number of penalties in x .

3. BOUNDARY MODEL

Boundary Maps. The pixel lattice is again S_I . Let S_B denote another regular lattice interspersed among the pixels (see Figure 3) and of dimension $(N-1) \times (N-1)$; these are the "boundary sites". We will associate $s = (i, j) \in S_B$ with the pixel $(i, j) \in S_I$ to the upper left of s .



$$N = 3$$

Pixel sites (o) and boundary sites (+)

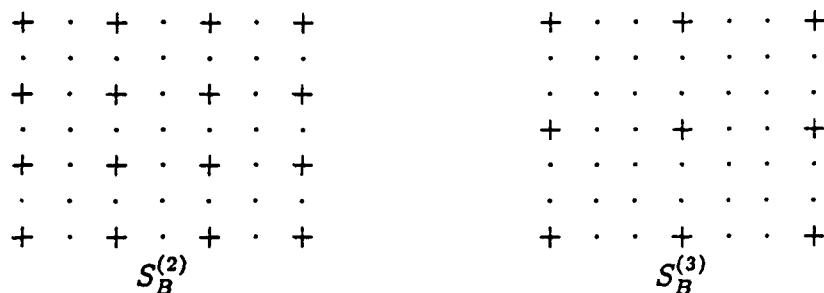
FIGURE 3

Given y , a grey-level image, we wish to assign values to the boundary variables $x = \{x_s, s \in S_B\}$, where $x_s = 1$ (resp. 0) indicates the presence (resp. absence) of a boundary at site $s \in S_B$. We have already discussed the corresponding interpretation of the boundary map $x = x(y)$ in terms of physical discontinuities in the underlying three-dimensional scene. We establish a boundary resolution or grid size $\sigma \geq 1$, analogous to the resolution used earlier for the partition model. Let $S_B^{(\sigma)} \subset S_B$ denote the sub-lattice $\{(i\sigma + 1, j\sigma + 1) : 1 \leq i, j \leq (N-2)/\sigma\}$. Only the variables $x_s, s \in S_B^{(\sigma)}$, interact directly with the data; the remaining variables $x_s, s \in S_B \setminus S_B^{(\sigma)}$, are determined by those on the "grid" $S_B^{(\sigma)}$. Figure 4 shows the grids $S_B^{(2)}$ and $S_B^{(3)}$ for $N=8$; the sites off the grid are denoted by dots. The selection of σ influences the interpretation of x , the computational load, the interaction range at the pixel level, and is related to the role played by the size of the spatial filter in edge detection methods based on differential operators. Finally, let Ω_I and $\Omega_B^{(\sigma)}$ denote the state spaces of intensity arrays and boundary maps respectively; that is,

$$\Omega_I = \{\{y_s\} : s \in S_I, 0 \leq y_s \leq 255\}, \quad \Omega_B^{(\sigma)} = \{\{x_s\} : s \in S_B^{(\sigma)}, x_s \in \{0, 1\}\}$$

Sometimes, we simply write Ω_B for $\Omega_B^{(1)}$.

Boundary-Data Interaction. Let $\langle s, t \rangle_\sigma, s, t \in S_B^{(\sigma)}$ denote a *nearest-neighbor* pair relative to the grid. Thus, $s = (i\sigma + 1, j\sigma + 1), t = (k\sigma + 1, l\sigma + 1)$ is such a (horizontal or vertical) pair if either $i = k$ and $j = l \pm 1$, or $j = l$ and $i = k \pm 1$. We identify $\langle s, t \rangle_\sigma$ with the elementary boundary segment consisting of the horizontal or vertical string of $\sigma + 1$ sites (in S_B) including s, t and the $\sigma - 1$ sites "in between".



Boundary grid at resolutions $\sigma = 2$ and $\sigma = 3$

FIGURE 4

The *disparity measure* should gauge the intensity "flux", $\Delta_{s,t} = \Delta_{s,t}^{(\sigma)}(y) \geq 0$, across $\langle s, t \rangle_\sigma$, i.e. orthogonal to the associated segment. We will experiment with several types of measures. An obvious choice at high resolution ($\sigma = 1$) is $\Delta_{s,t} = |y_{s^*} - y_{t^*}|$ where $s^*, t^* \in S_I$ are the two pixels associated with $\langle s, t \rangle_1$; see Figure 5.



Pixel pairs (o's) associated with horizontal and vertical boundary segments

FIGURE 5

The analogous choice at a lower resolution ($\sigma > 1$) might be a measure of the form $\Delta_{s,t} = m^{-1} |\sum_{D_{s^*}} y_t - \sum_{D_{t^*}} y_t|$, where $D_{s^*}, D_{t^*} \subset S_I$ are *adjacent blocks of pixels*, of the same size (m) and shape, and "separated" by $\langle s, t \rangle_\sigma$. These and other measures are discussed later on.

The energy function $U(x, y)$ should promote boundary maps x which are faithful to the data y in the sense that "large" values of $\Delta_{s,t}(y)$ are associated with "on" segments ($x_s, x_t = 1$) and "small" values with "off" segments ($x_s, x_t = 0$). There are no *a priori* constraints on x at this point; in fact, because of digitization effects, textures, and so-on, the energy U will typically favor maps x with undesirable dead-ends, multiple representations, high curvature, etc. These will be penalized later on. A simple choice for the x/y interaction is

$$(3.1) \quad U(x, y) = \sum_{\langle s, t \rangle_\sigma} (1 - x_s x_t) \phi(\alpha^{-1} \Delta_{s,t}(y))$$

where the summation extends over all nearest-neighbor pairs $\langle s, t \rangle_\sigma$; the "weighting function" $\phi(x)$, $x \geq 0$, and "normalizing constant" α will be described presently.

The energy in (3.1), which is similar to a "spin-glass" in statistical mechanics, is a variation of the ones we used in our previous work ([24],[25]); when $\sigma = 1$, the

variable x, x_t corresponds directly to the "edge" or "line" variables in [25] and [57]. Since y is given, the term $1 - x, x_t$ can be replaced by $-x, x_t$ with no change in the resulting boundary interpretation. By contrast, in [25] we were concerned with image *restoration* and regarded *both* x and y as unobservable; the data then consists of some transformation of y , involving for example blur and noise. In that case, or in conceiving U as defining a prior distribution over *both* y and x , the bond between the associated pixels should be broken when the edge is active, i.e. $1 - x, x_t = 0$. The term $1 - x, x_t$ is exactly analogous to the "controlled-continuity functions" in [68]. See also [43] for experiments involving simulations from a related Markov random field model; the resulting "fantasies" (y, x) are generated by stochastic relaxation and yield insight into the nature of these layered Markov models.

Returning to (3.1), a little reflection shows that ϕ should be increasing, with $\phi(0) < 0 < \phi(+\infty)$; otherwise, if ϕ were never negative, the energy would always be minimized with $x_s \equiv 1$. The intercept $\beta \doteq \phi^{-1}(0)$ is critical; values of Δ above (resp. below) the *threshold* $d^* \doteq \alpha\beta$ will promote (resp. inhibit) boundary formation. The influence of the threshold is reduced by choosing $\phi'(\beta) = 0$. We employ the simple quadratic

$$(3.2) \quad \phi(x) = \begin{cases} \left(\frac{x-\beta}{1-\beta}\right)^2, & x \geq \beta \\ -\left(\frac{x-\beta}{\beta}\right)^2, & 0 \leq x \leq \beta \end{cases}$$

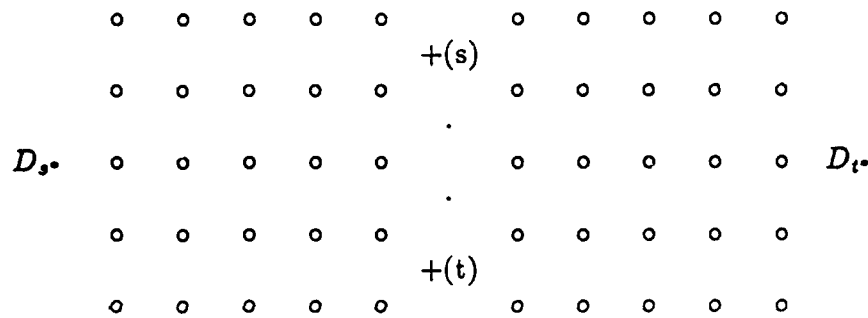
Notice that the maximum "penalty" ($\phi(0) = -1$) and "reward" ($\phi(1) = 1$) are balanced if we select $\alpha \approx \max \Delta_{s,t}$.

Disparity Measures. We employ one type of measure for depth and shape boundaries and another for the texture experiments. In the former case, the disparity measure involves the (raw) grey-levels only, whereas for texture discrimination we also consider data transforms based on the directional residuals (1.1). Except when $\sigma = 1$, the data sets are compared by the Kolmogorov-Smirnov distance.

At the highest resolution ($\sigma = 1$), the measure $|y_{s^*} - y_{t^*}|$ (where s^*, t^* are the two pixels associated with the boundary sites s, t - see Figure 5) can be effective for simple scenes but necessitates a single differential threshold $d^* = \alpha\beta$. Differences above d^* (resp. below d^*) promote (resp. inhibit) boundary formation. Typically, however, this measure will fluctuate considerably over the image, complicating the selection of d^* . (Such is the case, e.g. for the "cart" scene, see §5.) Moreover, this measure lacks any invariance properties, as will be explained below. A more effective measure is one of the form

$$(3.3) \quad \Delta_{s,t}(y) = \frac{|y_{s^*} - y_{t^*}|}{\gamma + \sum |y_{s_i^*} - y_{t_i^*}|}$$

where the sum extends over parallel edges $\langle s_i, t_i \rangle$ in the immediate vicinity of $\langle s, t \rangle$. Thus the difference $|y_{s^*} - y_{t^*}|$ is "modulated" by adjacent, competing differences. The result is a spatially-varying threshold and the distribution of $\Delta_{s,t}(y)$



Pixel blocks, D_{s*} and D_{t*} ,
associated with boundary segment $< s, t >_3$

FIGURE 6

across the image is less variable than that of $|y_{s*} - y_{t*}|$. Choosing $\gamma = \text{const.} \times \bar{\Delta}$, where $\bar{\Delta}$ is the mean (raw) absolute intensity difference over all (vertical and horizontal) bonds, renders $\Delta_{s,t}(y)$ invariant to linear transformations of the data; that is, $\Delta_{s,t}(y) = \Delta_{s,t}(ay + b)$ for any a, b .

At lower resolution, let D_{s*} and D_{t*} denote two adjacent blocks of pixels, of equal size and shape. An example is illustrated in Figure 6 for the case of two square blocks of size $5^2 = 25$ pixels which straddle a vertical boundary segment with $\sigma = 3$. Let $y(D_r) = \{y_s, s \in D_r\}$, $r = s^*, t^*$, be the corresponding grey-levels and set

$$(3.4) \quad \Delta_{s,t}(y) = d(y(D_{s*}), y(D_{t*})),$$

where d is the Kolmogorov-Smirnov distance discussed in §2. This is the disparity measure used for the House and Ice Floe scenes (see §5).

One difficulty with (3.4) is that the distance between two non-overlapping histograms is the maximum value, namely 1, regardless of the amount of separation. Thus, two constant regions differing by a single grey level are as "far apart" as two differing by 255 levels. Thus, it is occasionally necessary to "de-sensitize" (3.4), for example by "smearing" the data or perhaps adding some noise to it; see §5.

Raw grey-level data is generally not satisfactory for discriminating textures. Instead, as discussed in §2, we base the disparity measure on several data transformations, involving higher order spatial statistics, such as the directional residuals defined in (1.1). Given a family $\Lambda_1, \Lambda_2, \dots, \Lambda_m$ of these transforms (see §2), a resolution σ , and blocks D_{s*}, D_{t*} as above, define

$$(3.5) \quad \Delta_{s,t}(y) = \max_{1 \leq i \leq m} [c_i^{-1} d(y^{(i)}(D_{s*}), y^{(i)}(D_{t*}))]$$

where $y^{(i)} = \Lambda_i(y)$, $1 \leq i \leq m$, and $y^{(i)}(D_r) = \{y_s^{(i)}, s \in D_r\}$, exactly as in §2. Then $\Delta_{s,i}(y) > d^*$ (and, hence, $\phi(\alpha^{-1}\Delta_{s,i}(y)) > 0$) if and only if $d(y^{(i)}(D_{s^*}), y^{(i)}(D_{t^*})) \geq d^*c_i$ for some transform i . The thresholds c_1, \dots, c_m are again chosen to limit "false alarms". Finally, we note that (3.5) has the same desirable invariance properties as the measure constructed for partitioning (§2).

Penalties. $V(x)$ again denotes the total number of "penalties" associated with $x \in \Omega_B^{(\sigma)}$. These penalties are simply local binary patterns over subsets of $S_B^{(\sigma)}$. Figure 7 illustrates a family of four such patterns; they can be associated with any resolution σ by the obvious scaling.



Forbidden patterns

FIGURE 7

These correspond, respectively, to an isolated or abandoned segment, sharp turn, quadruple junction, and "small" structure. Depending on σ , the pixel resolution, and scene information, we may or may not wish to include the latter three. For example, including the last one with $\sigma = 6$ would prohibit detection of a square structure of pixel size 6×6 .

To further clarify the definition of V , identify each pattern, up to translation, with a set $C \subset S_B^{(\sigma)}$ and binary set $\xi = \{\xi_s, s \in C\}$. For instance, for $\sigma = 2$, the first pattern in Figure 7 is represented by $C = \{(1, 3), (3, 1), (3, 3), (3, 5)\}$ and corresponding ξ -values 0, 0, 1, 0. Let $\{C^{(j)}, \xi^{(j)}\}$, $1 \leq j \leq J$ be a family of patterns. Then

$$V(x) = \sum_{j=1}^J \sum_{\tau} \delta_{j,\tau}$$

where the inner sum extends over all translates τ and

$$\delta_{j,\tau} = \begin{cases} 1 & \text{if } x_{s+\tau} = \xi_s^{(j)} \text{ for all } s \in C^{(j)} \\ 0 & \text{otherwise} \end{cases}$$

Finally, there is a natural extension from $\Omega_B^{(\sigma)}$ to Ω_B which is useful for display and evaluation. Given $x \in \Omega_B^{(\sigma)}$, we define $x_s = 0$ for sites $s \in S_B \setminus S_B^{(\sigma)}$ lying on a row or column disjoint from $S_B^{(\sigma)}$, and $x_s = x_{t_1} x_{t_2}$ if s lies on a segment $\langle t_1, t_2 \rangle_\sigma$. Thus, for example, the state $x \in \Omega_B^{(2)}$ in Figure 8(a) is identified with the state

$x \in \Omega_B$ in 8(b).

$$\begin{array}{ccccc} 1 & \cdot & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 1 & \cdot & 1 \end{array}$$

8(a)

$$\begin{array}{ccccc} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{array}$$

8(b)

Completion of boundary configuration, from $\sigma = 2$ to $\sigma = 1$

Figure 8

Summary. We are given

- (i) a grey-level image $y = \{y_{ij}\}$;
- (ii) a resolution level $\sigma = 1, 2, \dots$;
- (iii) a disparity measure $\phi(\alpha^{-1}\Delta_{s,t}(y))$ for each neighbor pair $\langle s, t \rangle_\sigma$ in the sub-lattice $S_B^{(\nu)}$;
- (iv) a collection of penalty patterns.

The (MAP) boundary estimate $\hat{x} = \hat{x}(y)$ is any solution $x \in \Omega_B^{(\sigma)}$ of the constrained optimization

$$\text{minimize}_{x: V(x)=0} \sum_{\langle s, t \rangle_\sigma} (1 - x_s x_t) \phi(\alpha^{-1}\Delta_{s,t}(y))$$

where $V(x)$ is the number of penalties in x .

4. ALGORITHMS

We begin with an abstract formulation of the optimization and sampling problems outlined in Sections 1-3. We are actually interested in *posterior* distributions, $\Pi(x|y)$, but y is fixed by observation, and can be ignored in this discussion of computational issues. Thus, we are given two functions U and V on a space of configurations $\Omega = \{(x_{s_1}, x_{s_2}, \dots, x_{s_M}) : x_{s_i} \in \Lambda, 1 \leq i \leq M\}$ where Λ is finite, M is very large, and $S = \{s_1, s_2, \dots, s_M\}$ is a collection of "sites", typically a 2-D lattice. Write $x = (x_{s_1}, \dots, x_{s_M})$ for an element of Ω , and let

$$\Omega^* = \{x : V(x) = \underline{v}\}, \quad \underline{v} = \min_x V(x)$$

$$\Pi^*(x) = \delta_{\Omega^*}(x) \frac{\exp\{-U(x)\}}{\sum_{x' \in \Omega^*} \exp\{-U(x')\}}$$

We wish to solve the constrained optimization problem minimize $\{U(x) : V(x) = \underline{v}\}$ or to sample from the Gibbs distribution Π^* . (Recall that sampling allows us to entertain estimates other than the MAP.)

We have studied [25] Monte Carlo site-replacement algorithms for the unconstrained versions of these problems: stochastic relaxation (SR) for sampling, and stochastic relaxation with simulated annealing (SA) for optimization. SA was devised in [12] and [47] for minimizing a "cost functional" U (e.g. the tour length for the traveling salesman problem) by regarding U as the energy of a physical system and simulating the dynamics of chemical annealing. The effect is to drive the system towards the "ground states", i.e. the minimizers of U . This is accomplished by applying the Metropolis (relaxation) algorithm to the Boltzmann distribution

$$(4.1) \quad \frac{e^{-U(x)/t}}{\sum_{x'} e^{-U(x')/t}}$$

at successively lower values of the "temperature" t .

We presented two theorems in [25]: one for generating a sequence $\{X(k)\}$ which converges in distribution to (4.1) for $t = 1$ (SR), and one for generating a sequence $\{X(k)\}$ having asymptotic distribution the uniform measure over $\Omega_o = \{x \in \Omega : U(x) = \underline{u}\}$, $\underline{u} = \min_x U(x)$, (SR with SA). The essence of the latter algorithm is a "cooling schedule" $t = t_1, t_2, \dots$ for guaranteeing convergence. SA has been extensively studied recently ([5],[14],[28],[29],[35],[39],[41],[67]); see also the comprehensive review [1] and the references therein. Applications have emerged in neural networks and circuit design, to name but two areas.

Results concerning constrained SR and SA are reported in [23], which was motivated by a desire to find a theoretical foundation for the algorithms used here.

We have deviated from the instructions in [23], with regard to the cooling schedule, but at least we know that the algorithms represent approximations to rigorous results.

Both algorithms produce a *Markov chain* on Ω by sampling from the low-order, marginal conditional distributions of the free Gibbs measures

$$\Pi(x; t, \lambda) = \frac{\exp \{-t^{-1}(U(x) + \lambda V(x))\}}{\sum_{x'} \exp \{-t^{-1}(U(x') + \lambda V(x'))\}}$$

It is easy to check that

$$(4.2) \quad \lim_{\lambda \rightarrow \infty} \Pi(x; 1, \lambda) = \Pi^*(x)$$

and that

$$(4.3) \quad \lim_{\substack{\lambda \rightarrow \infty \\ t \rightarrow 0}} \Pi(x; t, \lambda) = \begin{cases} |\Omega_o^*|^{-1}, & x \in \Omega_o^* \\ 0, & \text{otherwise} \end{cases}$$

where $\Omega_o^* = \{\omega \in \Omega^* : U(\omega) = \xi\}$, $\xi = \min_{x \in \Omega^*} U(x)$. Let Π_o denote the uniform measure in (4.3). Sampling directly from $\Pi(x; t, \lambda)$ is impossible due to the size of Ω ; otherwise just use (4.2) and (4.3) to generate a sequence of random variables $X(k)$, $k = 1, 2, \dots$, with values in Ω , and limiting distribution either Π^* or Π_o . However, we can evaluate *ratios* $\Pi(x; t, \lambda)/\Pi(z; t, \lambda)$, $x, z \in \Omega$, and hence *conditional* probabilities. The price for *indirect* sampling is that we must restrict the rate of growth of λ and the rate of decrease of t .

Fix two sequences $\{t_k\}$, $\{\lambda_k\}$, a "site visitation" schedule $\{A_k\}$, $A_k \subset S$, and let $\Pi_k(x) = \Pi(x; t_k, \lambda_k)$. The set A_k is the cluster of sites to be updated at "time" k ; the "centers" of the clusters are addressed in a raster scan. In our experiments we take either $|A_k| \equiv 1$ or $|A_k| \equiv 5$, in which case the A_k 's are of the form $\{(i, j), (i+1, j), (i-1, j), (i, j+1), (i, j-1)\}$.

Define a non-homogeneous Markov chain $\{X(k), k = 0, 1, 2, \dots\}$ on Ω as follows. Put $X(0) = \eta$ arbitrarily. Given $X(k) = (X_{s_1}(k), \dots, X_{s_M}(k))$, define $X_s(k+1) = X_s(k)$ for $s \notin A_{k+1}$ and let $\{X_s(k+1) : s \in A_{k+1}\}$ be a (multivariate) sample from the conditional probability distribution $\Pi_{k+1}(x_s, s \in A_{k+1} | x_s = X_s(k), s \notin A_{k+1})$. Then, under suitable conditions on $\{t_k\}$ and $\{\lambda_k\}$, either

$$\lim_{k \rightarrow \infty} P(X(k) = x | X(0) = \eta) = \Pi^*(x)$$

or the limit is $\Pi_o(x)$. The condition in the former case (constrained SR) is that $t_k \equiv 1$, $\lambda_k \nearrow \infty$, and $\lambda_k \leq \text{const.} \cdot \log k$. The condition for convergence to Π_o (constrained SA) is that $t_k \searrow 0$, $\lambda_k \nearrow \infty$ and $t_k^{-1} \lambda_k \leq \text{const.} \cdot \log k$. The algorithm yields a solution to the constrained optimization problem (1.2) in the sense that the

asymptotic distribution of $X(k)$ is uniform over the solution set: if the solution is unique, i.e. $\Omega_o^* = \{x_o\}$, then $X(k) \rightarrow x_o$ in probability. See [23] for proofs.

Approximations. The logarithmic rate is certainly slow. Still, we often adhere to it for ordinary annealing; others ([44],[57]) have as well. We refer the reader to [43] for some interesting comparisons between schedules and to [67] for "fast annealing" algorithms. It is commonplace to find linear ($t_k = t_o - ak$) and exponential ($t_k = (1 - \gamma)^k t_o$, γ small) schedules; here k refers to the number of sweeps or iterations of S ; in our experiments $S = S_L^{(\sigma)}$ or $S_B^{(\sigma)}$.

We now describe several protocols used in our experiments. One variant we do *not* use is to fix $\lambda_k \equiv \lambda$ very large and do ordinary annealing, which might appear sensible since the solutions to $\min\{U(x) : V(x) = 0\}$ coincide with those of $\min\{U(x) + \lambda V(x)\}$ for all λ sufficiently large (due to the fact that Ω is finite). However this is not practical: unless t_o is very large and t_k is reduced very slowly, the system immediately gets stuck in local energy minima of $U + \lambda V$ which are basically independent of the data, although faithful to the constraints. *It is better to begin with states faithful to the data and slowly impose the constraints*, a standard technique in conventional optimization.

One variation of constrained SR that has been effective is "low-temperature sampling": fix $t_k \equiv \epsilon$ (small) and let $\lambda_k \nearrow \infty$. The idea is to reach a likely state of the posterior distribution $\Pi(x|y)$. In practice, we allow λ_k to grow *linearly*; the details are in Section 5.

Another variation is the analogue for constrained relaxation of "zero-temperature" sampling, which has been extensively studied by Besag [2] under the name ICM (for "iterated conditional modes"); see also [15], [18] and [27]. Without constraints, this algorithm, which is deterministic, results in a sequence of states $X(k)$ which monotonically decrease the global energy, i.e. increase the posterior likelihood. The constrained version operates as follows. Recall that when the set of sites A_{k+1} is visited for updating, we defined $X(k+1)$ by replacing the coordinates of $X(k)$ in A_{k+1} by a *sample* drawn from the conditional distribution of Π_{k+1} on $\{x_s, s \in A_{k+1}\}$ given the values $\{x_s = X_s(k), s \notin A_{k+1}\}$. Suppose we replace the sample with the *mode*, i.e. the most likely vector $\{x_s, s \in A_{k+1}\}$ conditional upon $\{x_s = X_s(k), s \notin A_{k+1}\}$. In essence, we fix $t_k \equiv 0$. This generates a *deterministic* sequence $X(k)$, $k = 0, 1, 2, \dots$ depending only on $X(0)$, Π_k , and $\{A_k\}$. (Notice that the mode is unaffected by t_k since it corresponds to the minimum of $U(x) + \lambda_k V(x)$.) Then, *during the k^{th} sweep*, with $\lambda \equiv \lambda_k$, the energy $U + \lambda_k V$ is successively reduced, just as in ICM where $\lambda_k \equiv 0$. Of course since there is no *fixed* (reference) energy, the algorithm cannot be conceived as one of iterative improvement. Several experiments were run with both the stochastic and deterministic algorithms; see Section 5.

5. EXPERIMENTS ¹

PARTITION MODEL

There are three experiments: an L-band synthetic aperture radar (SAR) image ² of ice floes in the ocean (Figure 9), a texture mosaic constructed from the Brodatz album [8] (Figure 10), and another mosaic from pieces of rug, plastic and cloth (Figure 11).

Processing. In each experiment the partitioning was randomly initiated; the labels, $x, s \in S_L^{(\sigma)}$, were chosen independently and uniformly from $0, 1, \dots, P-1$. Thereafter, label sites were visited and updated one at a time, by a "raster scan" sweep through the label array. MAP partitionings were approximated by "zero-temperature" sampling (see §4), with $\lambda = \lambda_k$ increasing with the number of sweeps. Specifically, λ was held at 0 through the first 10 sweeps, and thereafter was raised by 1 every 5 sweeps: $\lambda_k = 0, k = 1, \dots, 10$; $\lambda_k = 1, k = 11, \dots, 15$; $\lambda_k = 2, k = 16, \dots, 20$; etc. Most probably, λ could have been increased more rapidly, perhaps with every sweep, without substantially changing the results, but this was not systematically investigated. For the three experiments shown in Figures 9, 10, and 11, between 15 and 50 sweeps sufficed to bring the changes in labels to a halt; see below for more details. Recall that zero-temperature sampling corresponds to choosing the conditional mode. Occasionally there are ties, and these were resolved by choosing randomly, and uniformly, from the collection of modes.

As a general rule, results were less reliable at higher resolutions (lower σ 's) and when more labels were allowed (higher values of P). In these cases, repeated experiments, with different initializations, often produced different results. With P too large, homogeneous regions were frequently subdivided, being assigned two or three labels. With σ too small, the tendency was to mislabel small patches within a given texture. It is likely that many of these mistakes correspond to local minima; perhaps some could be corrected by following a proper annealing schedule (see §4), and by more careful choices of thresholds (see below). Here again, definitive experiments have not been done.

Measures of Disparity. Recall that the disparity measure is derived from the Kolmogorov-Smirnov distance between blocks of pixel data under various transformations, as defined in §2, equation (2.3). For the SAR image, good partitionings were obtained using only the raw data: $m = 1$ and $y_{(1)}$ is just y in equation (2.3).

¹ Fortran code and terminal sessions are available.

² We are grateful to the Radar Division at ERIM for providing us with the SAR image (collected for the U.S. Geological Survey under Contract 14-08-0001-21748 and the Office of Naval Research under Contract N-00014-81-C-0692 and N-00014-81-C-0295).

Evidently, grey-level distributions are enough to segment the water and ice "textures", at least when supplemented by the "prior constraints" embodied in the penalty term, $V(x)$.

The texture collages in Figures 10 and 11 are harder. We used four data transformations in addition to the raw pixel data. Hence, for these experiments $m = 5$, $y^{(1)} = y$, and $y^{(2)}, \dots, y^{(5)}$ are based on various transforms. In particular:

$y_s^{(2)}$ measures the intensity *range* in the 7×7 pixel block, V_s , centered at s :

$$y_s^{(2)} = \max_{t \in V_s} y_t - \min_{t \in V_s} y_t$$

$y_s^{(3)}$ is the "residual" (equation (1.1)) obtained by comparing y_s to the 24 "boundary pixels" (∂V_s) of V_s (i.e. all pixels on the perimeter of the 7×7 block):

$$y_s^{(3)} = |y_s - \frac{1}{24} \sum_{t \in \partial V_s} y_t|$$

and $y_s^{(4)}$ and $y_s^{(5)}$ are horizontal and vertical "directional residuals":

$$y_s^{(4)} = |y_s - \frac{1}{2}(y_{s+(1,0)} + y_{s+(-1,0)})|$$

$$y_s^{(5)} = |y_s - \frac{1}{2}(y_{s+(0,1)} + y_{s+(0,-1)})|$$

Parameter Selection. The resolution (σ) was 7 for the SAR picture (Figure 9); 15 for the Brodatz collage; and 13 for the pieces of rug, plastic and cloth. These numbers were chosen more or less ad hoc, but are small enough to capture the important detail of the respective pictures while not so small as to incur the degraded performance, seen at higher resolutions and mentioned earlier.

The number of allowed labels is also important; recall that too many usually results in over-segmentation. This was actually used to advantage in the SAR experiment (Figure 9), where there are evidently two varieties of ice. The best segmentations were obtained by allowing three labels. Invariably, two would be assigned to the ice, and one to the water. Using just two labels led to mistakes within the ice regions, although there was little experimentation with the Kolmogorov-Smirnov threshold, and no attempt was made with the data transforms ($m > 1$) used for the collages. In the other experiments, the number of labels was set to the number of texture species in the scene.

The most important parameters were the thresholds, $\{c_i\}$ $1 \leq i \leq m$, associated with the Kolmogorov-Smirnov statistics (see (2.3)). For the SAR experiment, $m =$

1, and the threshold was guessed, a priori; it was found that small changes are reflected only in the lesser details of the segmentation. For the collages ($m = 5$), the thresholds were chosen by examining histograms of Kolmogorov-Smirnov distances for block pairs *within homogeneous samples of the textures*. Thresholds were set so that no more than three or four percent of these intra-region distances would be above threshold (a "false alarm"). Of course, we would have preferred to find more or less universal thresholds, one for each data transform, but this may not be possible. Conceivably, with enough of the "right" transforms, one could set conservative (high) and nearly universal thresholds, and be assured that visibly distinct textures would be segmented with respect to at least one of the transforms. Recall that the disparity measure (equation (2.3)) is constructed to signal "different" when the distance between blocks, with respect to *any* of the transforms, exceeds threshold.

Figure 9 (SAR). As mentioned earlier, three labels were used, with the expectation that the ice would segment into two regions (basically, dark and light). The resolution was $\sigma = 7$, and the Kolmogorov-Smirnov statistic was computed only on the raw data, so $m = 1$. The threshold was $c_1 = .15$. The original image is 512×512 (the pixel resolution is about 4m by 4m), but to avoid special treatment of the boundary, only the 462×462 piece shown in 9(A) was processed. The label lattice, $S_L^{(\sigma)}$, is 64×64 . Figure 9(B) shows the evolution of the partitioning during the relaxation. For display, grey levels were arbitrarily assigned to the labels. The upper left panel is the random starting configuration. In successive panels are the states of the labels after each five iterations (full sweeps). In the bottom right panel, the two labels associated with ice are combined, "by hand".

Figure 10 (Brodatz Textures). The Kolmogorov-Smirnov thresholds were $c_1 = .40$, $c_2 = .53$, $c_3 = .26$, $c_4 = .28$, and $c_5 = .19$, corresponding to the transforms $y^{(1)}, \dots, y^{(5)}$ discussed above. A 246×246 piece of the original 256×256 image was processed, and is shown in Figure 10(A). Leather and water are on top, grass and wood on the bottom, and sand is in the middle. The resolution was $\sigma = 15$, which resulted in a 16×16 label lattice $S_L^{(\sigma)}$. Figure 10(B) shows the random starting configuration (upper left panel), the configuration after 5 iterations (upper right panel), after 10 iterations (lower left panel), and after 15 iterations (lower right panel), by which point the labels had stopped changing.

Figure 11 (Rug, Plastic, Cloth). The 216×216 image in Figure 11(A) was partitioned at resolution $\sigma = 13$, with a 16×16 label lattice. The Kolmogorov-Smirnov thresholds were $c_1 = .90$, $c_2 = .49$, $c_3 = .20$, $c_4 = .11$, and $c_5 = .12$, corresponding to the same data transforms used for the Brodatz textures (Figure 10). The experiment makes apparent a *hazard* of long range bonds: the gradual but marked lighting variation across the top of the image produces a large Kolmogorov-Smirnov distance when raw pixel blocks from the left and right sides are compared. This makes it necessary to essentially ignore the raw data Kolmogorov-Smirnov

statistic, and base the partitioning on the four data transformations; hence the threshold $c_1 = .9$. The transformed data are far less sensitive to lighting gradients. Figure 11(B) displays the evolution of the partitioning during relaxation. The layout is the same one used in the previous figures, showing every 5 iterations, except that there are 10 iterations between the final two panels. The lower right panel is the partitioning after the 30'th sweep, by which time the pattern was "frozen".

PLACE FIGURES 9, 10, 11 HERE

BOUNDARY MODEL

There are five test images: one made indoors from tinkertoys ("cart"), an outdoor scene of a house, another of ice floes in the ocean (the same SAR image used above), and two texture mosaics constructed from the Brodatz album.

Processing. All the experiments were performed with the same site-visitation schedule. Given the resolution σ , which varies among experiments, the sites of the sub-lattice $S_B^{(\sigma)}$ were addressed in a raster-scan and *five sites* were simultaneously updated. Specifically, at each visit to the site $(i\sigma + 1, j\sigma + 1)$, the values of the boundary process at this site and its four nearest neighbors, $\{((i \pm 1)\sigma + 1, (j \pm 1)\sigma + 1)\}$, were replaced based on the conditional distribution of these five boundary variables given the variables at the other sites and the data y . Of course this distribution is concentrated on the $2^5 = 32$ possible configurations for these five variables.

Two update mechanisms were employed: stochastic relaxation and the "zero-temperature", deterministic variation discussed earlier. In the former case, the updated binary quintuple is a sample from the aforementioned conditional distribution, which varies depending on the penalty weight λ_k for the k 'th sweep of the lattice $S_B^{(\sigma)}$. Of course "0-temperature" refers to replacing the sample by the conditional mode.

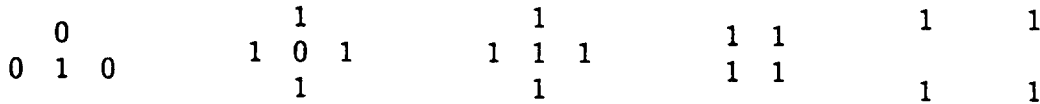
Constrained simulated annealing was not used, at least not in accordance with the formula in which $t_k \searrow 0$, $\lambda_k \nearrow \infty$ and $t_k^{-1} \lambda_k \leq \text{const.} \cdot \log k$ (k = sweep number). Instead, we let λ_k grow *linearly* and, in the case of stochastic relaxation, we *fixed* the temperature t_k at some "small" value.

Stochastic relaxation at low temperature is more effective than at zero temperature (essentially iterative improvement). However, deterministic relaxation sufficed

for all but two scenes, the ice floes and the four texture collage; these results could not be duplicated with deterministic relaxation. In one case, we present both results for comparison.

Generally, deterministic relaxation stabilizes in 5 to 10 sweeps whereas stochastic relaxation requires more sweeps, perhaps 20 to 60. We provide several pictures showing the evolution of the algorithm.

Penalties. All the experiments were conducted with the same forbidden patterns, namely those in Figure 12, with the exception of the house scene, for which the last pattern was omitted. (At the resolution used for the house, namely $\sigma = 3$, the inclusion of that pattern would inhibit the formation of structures at the scale of six pixels; many such non-trivial structures appear in that scene.) Thus, the penalty function $V(x)$ records a unit penalty for each occurrence in the boundary map $x = \{x_s, s \in S_B^{(\sigma)}\}$ of any of the five patterns depicted in Figure 12. It is interesting to note that in no case was the final labelling completely free of penalties, i.e. $V(\hat{x}) = 0$. Perhaps this could be achieved with a proper annealing schedule, or with updates of more than five sites.



Forbidden patterns

FIGURE 12

Measures of Disparity. All the experiments are based on instances of the measures (3.3)-(3.5) described in §3.

(i) For the first experiment, the cart scene, the boundary resolution is $\sigma = 1$ and we employed the measure given in (3.3) with $\gamma = 10\bar{\Delta}$ and the raw difference $|y_{s^*} - y_{t^*}|$ modulated by the four nearest differences of the same orientation as $\langle s^*, t^* \rangle$. Thus, for the horizontal pair $\langle s, t \rangle$ of adjacent boundary sites,

$$\Delta_{s,t}(y) = \frac{|y_{s^*} - y_{t^*}|}{10\bar{\Delta} + \sum_{k=\pm 1, \pm 2} |y_{i,j+k} - y_{i+1,j+k}|}$$

where $s^* = (i, j)$, $t^* = (i + 1, j)$, and $\bar{\Delta}$ is the mean absolute intensity difference over the image. The utility seems largely impervious to the choice of the scaling constant (here = 10) for the mean as well as to the range of the modulation.

(ii) We used the Kolmogorov-Smirnov measure (3.4) for both the house and ice floes scenes. For the house, we chose $\sigma = 3$ and blocks of size 25; the set-up is depicted in Figure 6. Due to the uniform character of the background (e.g. the

sky) the distance (3.4) was computed based on the transformed data $y'_{ij} = y_{ij} + \eta_{ij}$, where $\{\eta_{ij}\}$ are independent variables, and distributed with a triangular density, specifically that of $10(u_1 + u_2)$, u_1, u_2 uniform (and independent) on $[0, 1]$.

The boundary resolution for the radar experiment is $\sigma = 8$, reflecting the larger important structures there; the image is 512×512 . The dynamic range is very narrow and the difference between the dark water and somewhat less dark ice is essentially one of texture, due in part to the customary speckle noise. In particular, the ice cannot be well-differentiated from the water based on shading alone. The disparity measure is (3.4), applied to the raw image data over 24×24 blocks. The problem encountered in the house scene is actually *alleviated* by the speckle.

(iii) The texture mosaic experiments are based on the measure (3.5) for a particular family $\Lambda_1, \dots, \Lambda_5$ of five data transformations or "features". In each case, the resolution is $\sigma = 5$ and block size is 21×21 . Recall that these five features are combined into a single measure of change according to the formula in (3.5). The transformations used are the *range*

$$z_s^{(1)} = \max_{t \in V_s} y_t - \min_{t \in V_s} y_t$$

over a 7×7 window V_s , centered at pixel s , and the four *directional residuals*

$$z_s^{(2)} = |y_s - \frac{1}{2}(y_{s+(0,1)} + y_{s+(0,-1)})|$$

$$z_s^{(3)} = |y_s - \frac{1}{2}(y_{s+(1,0)} + y_{s+(-1,0)})|$$

$$z_s^{(4)} = |y_s - \frac{1}{2}(y_{s+(1,1)} + y_{s+(-1,-1)})|$$

$$z_s^{(5)} = |y_s - \frac{1}{2}(y_{s+(-1,1)} + y_{s+(1,-1)})|.$$

These residuals were then uniformly averaged over V_s , yielding the final features $y^{(1)}, \dots, y^{(5)}$.

It is instructive to compare the Kolmogorov-Smirnov differences for the raw and transformed data over these texture mosaics. Typically, if one looks at the resulting two histograms of differences for a give transform, one finds that, whereas the raw (Kolmogorov-Smirnov) differences are actually larger at the texture borders, the *transitions* between the borders and interiors are sharper for the transformed data. Detecting the boundaries with the raw data necessitates an unacceptable number of "false alarms" in the sense of interior "micro-edges".

Finally, the values of the constants c_1, \dots, c_5 used in the construction of $\Delta_{s,t}$ (see (3.5)) are selected by restricting the percentage of false alarms. The details are given in the following section.

Parameter Selection. Recall that the total change across the boundary segment $\langle s, t \rangle$ is measured by $\phi(\alpha^{-1}\Delta_{s,t}(y))$, where ϕ is given in (3.2). Given Δ , there are two parameters to choose: a normalizing constant α and the intercept $\beta = \phi^{-1}(0)$; the "threshold" for Δ is then $d^* \doteq \alpha\beta$.

For the object boundary experiments, namely the cart, house and ice floes, the parameters α and β were chosen as follows. Find the *mean* disparity over all (vertical and horizontal) values of $\Delta_{s,t}$ for relevant bonds $\langle s, t \rangle$; take α equal to the 99th percentile of those above the mean and d^* equal to the 70th percentile of those above the mean. This yields the values $\alpha = 150$, $\beta = .28$ for the cart scene; recall that for this experiment, both the grid and block sizes are unity. For the house scene ($\sigma = 3$) the Kolmogorov-Smirnov statistics were computed over 5×5 blocks, and the resulting parameters are then $\alpha = 1$ and $\beta = .7$. (The number of distances at (the maximum) value $\Delta = 1$ was considerable.) Finally, for the ice floes, the recipe above yielded $\alpha = .33$, $\beta = .40$.

Turning to the experiments with texture mosaics, let $c_{i,k}$ denote the normalizing constant in (3.5) for feature i , $1 \leq i \leq 5$, and texture k , $1 \leq k \leq K$, where K is the number of textures in the mosaic. For each feature i and texture type k , we computed the histogram of the (combined vertical and horizontal) Kolmogorov-Smirnov distances and selected $c_{i,k} = 100(1 - \gamma)$ percentile of that histogram. Specifically, we took $\gamma = .01$ for the two Brodatz collages. (Other experiments indicated that *any* (small) value of γ will suffice, say $0 \leq \gamma \leq .03$.) Thus, $100(1 - \gamma)$ percent of the distances $d(y^{(i)}(D_1), y^{(i)}(D_2))$ are below $c_{i,k}$ within each texture type k . Now set $c_i = \max_{1 \leq k \leq K} c_{i,k}$, insuring that at most $k\gamma(100)$ percent of the *interior* differences $\Delta_{s,t}(y)$ within the entire collage will exceed the threshold $d^* = 1$. Finally, since α and β are then constrained by $\alpha\beta = 1$, we put $\alpha = 2$ and $\beta = .5$.

Figure 13 (Cart Scene). Sixty sweeps of stochastic relaxation were run with $t_k \doteq .05$ and $\lambda_k \nearrow 3$. Actually, all the boundaries were "in place" after about 10 sweeps, as illustrated in Figure 13(B), which shows every third sweep up to the forty-sixth. Figure 13(C) shows the forty-sixth sweep at larger scale. The image is 110×110 . Not shown is a run with the deterministic algorithm; the results are virtually indistinguishable.

Figure 14 (House Scene). This 256×256 monochrome image was supplied to us by the VISIONS group at the University of Massachusetts. The update is by deterministic relaxation with λ_k increasing linearly from $\lambda_0 = 0$ to $\lambda_{10} = 2$.

Figure 15 (Ice Floes). The image (15(A)) is 512×512 . We did sixty sweeps of stochastic relaxation with $t_k \doteq .1$ and $\lambda_k \nearrow 2$. Figure 15(B) shows sixteen "snapshots" - every third sweep as in Figure 13, as well as the final (60th) sweep.

Figure 16 (Brodatz Collage 1). The collage is composed of nine Brodatz textures (16(A)): leather, grass, and pigskin (top row), raffia, wool, and straw (mid-

er, wood, and sand (bottom row). Two of the textures, leather and
 ed in the two circles. The image size is 584×384 , the individual
 128×128 . We show the results (16(B)) of both the deterministic
 tic (right) algorithms; they are roughly comparable. Other false
 .005, .02, and .03) yield the same overall quality.

rodatz Collage 2). There are four textures (17(A)): raffia (upper
 right), wool (bottom), and pigskin (center). Two runs (different
 (17(B)), individual frames representing every third sweep.

PLACE FIGURES 13 THROUGH 17 HERE

ibined into a single
 izations.

dary unit or primi-
 dependent, namely
 is detected and lo-
 parity measure has
 $< s, t >_\sigma$ by other
 six represented (up

+
 +
 +

tion term

In Figure 18, these
 . One can imagine
 ls to straddle (and

veral sensors, e.g.
 ar. One might in-
 used in the texture

especially if range
 ess values, and let
 ccluding (=depth)
 low). Now rig the
 and $\Delta(y^{(2)})$ with
 sponding terms to
 d-ends, etc. (with
 iding to physically

egion labels into a
 penalize improper

local configurations in the pair (x^b, x') , for example "type 1" errors (a boundary "between" like region labels) and "type 2" errors (no boundary "between" unlike labels). The problem may be that there are deep local minima which are unfaithful to the data but difficult to escape from, at least without updating many sites.

7. SUMMARY

We have developed algorithms for partitioning an image, possibly textured, into homogeneous regions and for locating boundaries at significant transitions. Both are based on a scale-dependent notion of disparity, or gradient, and both employ a Bayesian framework to make use of prior beliefs about regular boundary or region configurations.

The disparity measure scores the difference between the statistical structures of two scale-dependent blocks of pixels. We have experimented with several measures. Ideally, the disparity will be large when there is an apparent difference, either in grey-level or in texture, between the blocks. Usually, it was necessary to tune the measure to the particular textures or structures involved; a more universal measure may require both better preprocessing (e.g. first extracting reflectance from intensity [40]) and better use of "high-level" information about expected macrostructures and shapes. For texture discrimination, by either partitions or boundary placement, we introduce a class of features, or transformations, that are decidedly multivariate, depending on the spatial distribution of large numbers of pixel grey levels. Our disparity measure is then a composite of measures of differences in the histograms of the block data, under the various transformations. Low-order features, such as those derived solely from raw grey-level histograms and co-occurrence matrices, were not as effective in our framework.

Disparity measures between pairs of pixel blocks drive the segmentations or boundary placements through a "label model", that specifies likely label configurations conditional on disparity data. For partitioning, labels are generic and associated with local blocks of the image. Two labels are the same if their respective regions are judged to be instances of the same texture. For boundary placement, the labels are zero or one, and interpreted as indicating, respectively, the absence or presence of boundary elements. *A priori* knowledge about acceptable label configurations, which, for example, may preclude very small or thin regions, or cluttered boundary elements, is applied by restricting labels to an appropriate subset of all possible configurations. The result of modelling disparity-label interactions and of defining restricted configurations, is a Gibbs distribution jointly on pixel grey levels and label configurations, with the marginal label distribution supported on a subset of the configuration space.

Partitioning and boundary finding is accomplished by approximating the maximum a posteriori (MAP) label configuration, conditioned on observed pixel data. Because certain configurations are forbidden, MAP estimation amounts to constrained optimization. Stochastic relaxation and simulated annealing are extended to accommodate constraints by introducing a non-negative constraint function that is zero only for allowed label configurations. The constraint function, with a multiplicative constant, is added to the posterior energy, and the constant is slowly increased during relaxation. Straightforward calculations establish an upper bound

on the rate of increase of this multiplicative constant that insures convergence of the relaxation and annealing algorithms to the desired limits. In a series of partitioning and boundary-finding experiments, deterministic and other fast variations of the constrained relaxation algorithm are found to be effective.

The partitioning model is appropriate when a small number of homogeneous regions are present. Disjoint instances of a common texture are automatically identified. The boundary model can be effective in complex, multi-textured, scenes. Both models sometimes require prior training to adjust parameters.

ACKNOWLEDGEMENT

We are indebted to Elie Bienenstock for recommending the random topology employed in the partition model, and for pointing out a fascinating connection to some work in neural modelling [49].

REFERENCES

1. E. Arts and P. van Laarhoven, "Simulated annealing: a pedestrian review of the theory and some applications," in *NATO Advanced Study Institute on Pattern Recognition: Theory and Applications*, Spa, Belgium, 1986.
2. J. Besag, "On the statistical analysis of dirty pictures," (with discussion), *J. Royal Statist. Soc.*, series B, 48, pp. 259-302, 1986.
3. A. Blake, "The least disturbance principle and weak constraints," *Pattern Recognition Letters*, 1, pp. 393-399, 1983.
4. A. Blake and A. Zisserman, "Weak continuity constraints in computer vision," Report CSR-197-86, Dept. of Comp. Sci., Edinburgh Univ., 1986.
5. E. Bonomi and J.-L. Lutton, "The n-city travelling salesman problem: Statistical mechanics and Metropolis algorithm," *SIAM Review*, 26, pp. 551-568, 1984.
6. A. Brandt, "Multi-level approaches to large scale problems," in *Proceedings of the International Congress of Mathematicians 1986*. Ed. A. M. Gleason, American Mathematical Society, Providence, 1987.
7. J.S. Bridle and R.K. Moore, "Boltzmann machines for speech pattern processing," *Proc. Inst. of Acoustics, Windermere*, pp. 1-8, 1984.
8. P. Brodatz, *Texture: A Photographic Album for Artists and Designers*. New York: Dover, 1966.
9. J.B. Burns, A.R. Hanson, and E.M. Riseman, "Extracting straight lines," *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 425-455, 1986.
10. T.M. Cannon, H.J. Trussell, and B.R. Hunt, "Comparison of image restoration methods," *Applied Optics*, 17, pp. 3385-3390, 1978.
11. J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 679-698, 1986.
12. V. Cerny, "A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm," *Inst. Phys. and Biophysics, Comenius Univ., Bratislava*, 1982.
13. B. Chalmond, "Image restoration using an estimated Markov model," Dept. of Mathematics, University of Paris, Orsay, 1987.
14. T.-S. Chiang and Y. Chow, "On eigenvalues and optimal annealing rate," Institute of Mathematics, Academia Sinica, Taipei, Taiwan, preprint 1987.
15. F.S. Cohen and D.B. Cooper, "Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, 9, pp. 195-219, 1987.

16. G.R. Cross and A.K. Jain, "Markov random field texture models," *IEEE Trans. Pattern Anal. Machine Intell.*, 5, pp. 25-40, 1983.
17. H. Derin and H. Elliott, "Modelling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, 9, pp. 39-55, 1987.
18. H. Derin and C. Won, "A parallel image segmentation algorithm using relaxation with varying neighborhoods and its mapping to array processors," Tech. Report, Dept. of Elec. and Comp. Engrg., University of Massachusetts, 1986.
19. P.A. Devijver and M.M. Dekesel, "Learning the parameters of a hidden Markov random field image model: a simple example," in *Pattern Recognition Theory and Applications*, P.A. Devijver and J. Kittler, Eds., Heidelberg: Springer-Verlag, pp. 141-163, 1987.
20. A. Gagalowicz and Song De Ma, "Sequential synthesis of natural textures," *CVGIP*, 30, pp. 289-315, 1985.
21. D. Geman, "A stochastic model for boundary detection," *Image and Vision Computing*, May, 1987.
22. D. Geman, "Parameter estimation for Markov random fields with hidden variables and experiments with the EM algorithm," *Complex Systems Technical Report No. 21*, Div. of Applied Mathematics, Brown University, 1984.
23. D. Geman and S. Geman, "Relaxation and annealing with constraints," *Complex Systems Technical Report No. 35*, Div. of Applied Mathematics, Brown University, 1987.
24. D. Geman, S. Geman, and C. Graffigne, "Locating texture and object boundaries," in *Pattern Recognition Theory and Applications*, P.A. Devijver and J. Kittler, Eds., Heidelberg: Springer-Verlag, 1987.
25. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern anal. Machine Intell.*, 6, pp. 721-741, 1984.
26. S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," in: *Proceedings of the International Congress of Mathematicians, 1986*. Ed. A.M. Gleason, American Mathematical Society, Providence, 1987.
27. S. Geman and D.E. McClure, "Statistical methods for tomographic image reconstruction," in: *Proceedings of the 46th Session of the International Statistical Institute*, Bulletin of the ISI, Vol. 52, 1987.
28. B.Gidas, "Nonstationary Markov chains and convergence of the annealing algorithm," *J. Stat. Phys.*, 39, pp. 73-131, 1985.

29. B. Gidas, "A renormalization group approach to image processing problems," *IEEE Trans. Pattern Anal. Machine Intell.*, (to appear, 1989).
30. C. Graffigne, "Experiments in texture analysis and segmentation," Ph.D. Dissertation, Div. of Applied Mathematics, Brown University, 1987.
31. P.J. Green, Discussion: "On the statistical analysis of dirty pictures" by Julian Besag, *J. of the Royal Stat. Society*, B-48, pp. 259-302, 1986.
32. D.M. Greig, B.T. Porteous, and A.H. Seheult, Discussion: "On the statistical analysis of dirty pictures" by Julian Besag, *J. of the Royal Stat. Society*, B-48, pp. 259-302, 1986.
33. U. Grenander, "Tutorial in Pattern Theory," Lecture Notes Volume, Div. of Applied Mathematics, Brown University, 1984.
34. W.E.L. Grimson and T. Pavlidis, "Discontinuity detection for visual surface reconstruction," *CVGIP*, 30, pp. 316-330, 1985.
35. B. Hajek, "A Tutorial Survey of Theory and Applications of Simulated Annealing," *Proceedings of the 24th IEEE Conference on Decision and Control*, pp. 755-760, 1985.
36. R.M. Haralick, "Digital step edges from zero crossings of second directional derivatives," *IEEE Trans. Pattern Anal. Machine Intell.*, 6, pp. 58-68, 1984.
37. R.M. Haralick, K. Shanmugam, and I. Denstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybernetics*, 6, pp. 610-621, 1973.
38. G.E. Hinton and T.J. Sejnowski, "Optimal perceptual inference," *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 1983.
39. R. Holley and D. Stroock, "Simulated annealing via Sobolev inequalities," preprint, 1987.
40. B.K.P. Horn and M.J. Brooks, "The variational approach to shape from shading," *Comput. Vision Graphics Image Process.*, 33, pp. 174-208, 1986.
41. C.-R. Hwang and S.-J. Sheu, "Large time behaviors of perturbed diffusion Markov processes with applications," I, II, and III. Institute of Mathematics, Academia Sinica, Taipei, Taiwan, preprint 1987.
42. B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, 290, 1981.
43. A. Kashko and B.F. Buxton, "Markov random fantasies," preprint, 1986.
44. A. Kashko, H. Buxton, and B.F. Buxton, "Parallel stochastic optimization in computer vision," preprint, 1987.
45. R.L. Kashyap, R. Chellappa, and A. Khotanzed, "Texture classification using

features derived from random field models," *Pattern Recognition Lett.*, 1, pp. 43-50, 1982.

46. H.T. Kilver and N.A. Campbell, "Allocation of remotely sensed data using Markov models for spectral variables and pixel labels," preprint, 1986.

47. S. Kirkpatrick, C.D. Gellatt, Jr. and M.P. Vecchi, "Optimization by simulated annealing," *Science*, 220, pp. 671-680, 1983.

48. K. Laws, "Textured image segmentation," Ph.D. Dissertation, University of Southern California, Los Angeles, USCPI Rept. 940, 1980.

49. C. von der Malsburg and E. Bienenstock, "Statistical coding and short-term synaptic plasticity: a scheme for knowledge representation in the brain," in *Disordered Systems and Biological Organization*. Eds. E. Bienenstock, F. Fajelman Soulié, and G. Weisbuch, NATO ASI Series, Springer-Verlag Berlin, 1986.

50. D. Marr and E. Hildreth, "Theory of edge detectors," *Proc. Royal Soc., B*, 207, pp. 187-207, 1980.

51. J.L. Marroquin, "Surface reconstruction preserving discontinuities," *Artificial Intell. Lab. Memo* 792, M.I.T., 1984.

52. J.L. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Am. Stat. Assoc.*, Vol. 82, pp. 76-89, 1987.

53. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, Vol. 21, pp. 1087-1091, 1953.

54. A. Mitiche and J.K. Aggarwal, "Image segmentation by conventional and information integrating techniques: a synopsis," *Image and Vision Computing*, 3, pp. 50-62, 1985.

55. J. Moussouris, "Gibbs and Markov random systems with constraints," *J. Stat. Physics*, 10, pp. 11-33, 1974.

56. D. Mumford and J. Shah, "Boundary detection by minimizing functionals, I," preprint, 1986.

57. D.W. Murray and B.F. Buxton, "Scene segmentation from visual motion using global optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, 9, pp. 220-228, 1987.

58. D.W. Murray, A. Kashko, and H. Buxton, "A parallel approach to the picture restoration algorithm of Geman and Geman on an SIMD machine," *Image and Vision Computing*, 4, pp. 133-142, 1986.

59. V.S. Nalwa and T.O. Binford, "On detecting edges," *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 699-714, 1986.

60. E. Oja, "Texture subspaces," *Pattern Recognition Theory and Applications*, P.A. Devijver and J. Kittler, Eds., Heidelberg: Springer-Verlag, 1987.
61. T. Pavlidis, "A critical survey of image analysis methods," Expanded version of paper delivered to 8th International Conference on Pattern Recognition, Paris, October, 1986.
62. T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory," *Nature*, 317, pp. 314-319, 1985.
63. B.D. Ripley, "Statistics, images, and pattern recognition," *Canadian J. of Statistics*, 14, pp. 83-111, 1986.
64. A. Rosenfeld and L.S. Davis, "Image segmentation and image models," *Proc. IEEE*, 67, pp. 764-772, 1979.
65. B.W. Silverman. Discussion: "On the statistical analysis of dirty pictures," by Julian Besag, *J. of the Royal Statistical Society, Series B*, vol. 48, pp. 259-302, 1986.
66. R.H. Swendsen and J.-S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Physical Review Letters*, 58, pp. 86-88, 1987.
67. H. Szu, "Non-convex optimization," SPIE Conference on Real Time Signal Processing IX, 698, 1986.
68. D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 413-424, 1986.
69. V. Torre and T. Poggio, "On edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 147-163, 1986.
70. L. Van Gool, P. Dewaelle, and A. Oosterlinck, "Texture analysis anno 1983," *CVGIP*, 29, pp. 336-357, 1985.
71. F.M. Vilnrotter, R. Nevatia and K.E. Price, "Structural analysis of natural textures," *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 76-89, 1986.
72. H.L. Voorhees, Jr., "Finding texture boundaries in images. Masters Dissertation, Dept. of Electrical Engineering and Computer Science, Massachusetts Inst. of Technology, 1987.
73. W.A. Yasnoff, J.K. Mui and J.W. Bacus, "Error measures for scene segmentation," *Pattern Recognition*, 9, pp. 217-231, 1977.
74. L. Younes, "Estimation and annealing for Gibbsian fields," Tech. Rept., University of Paris, Dept. of Mathematics, Orsay, 1987.



Figure 9(A)

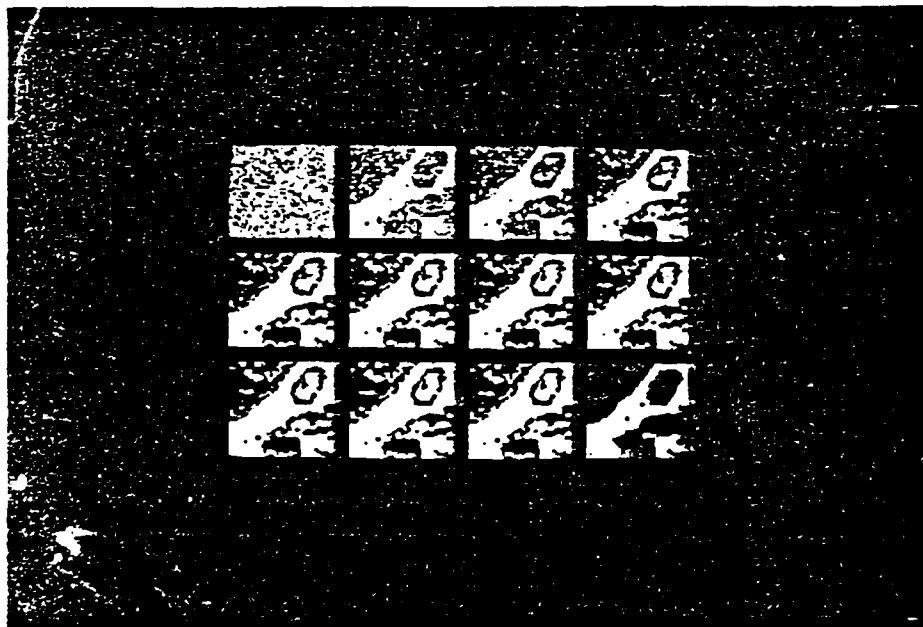


Figure 9(B)

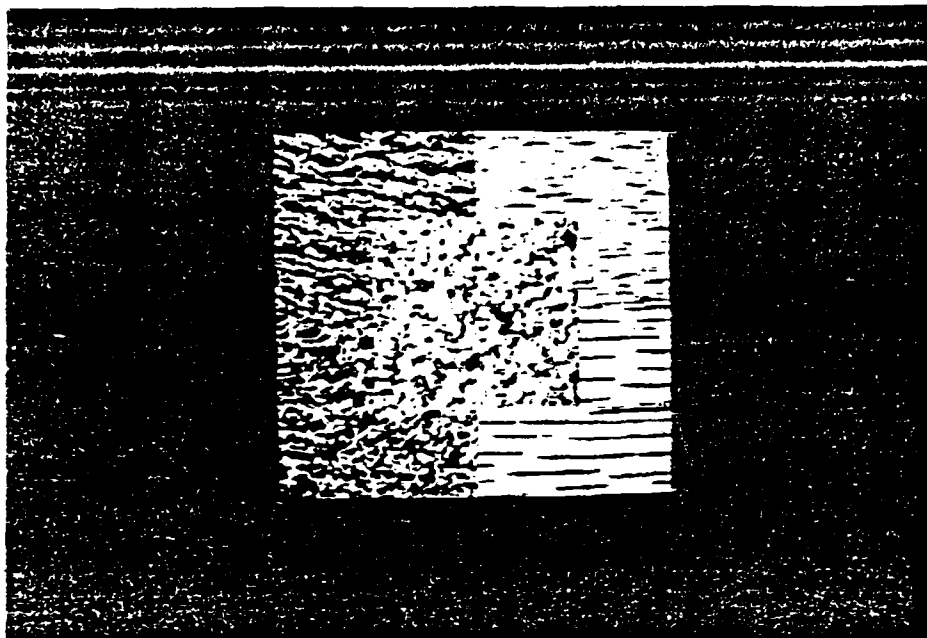


Figure 10(A)

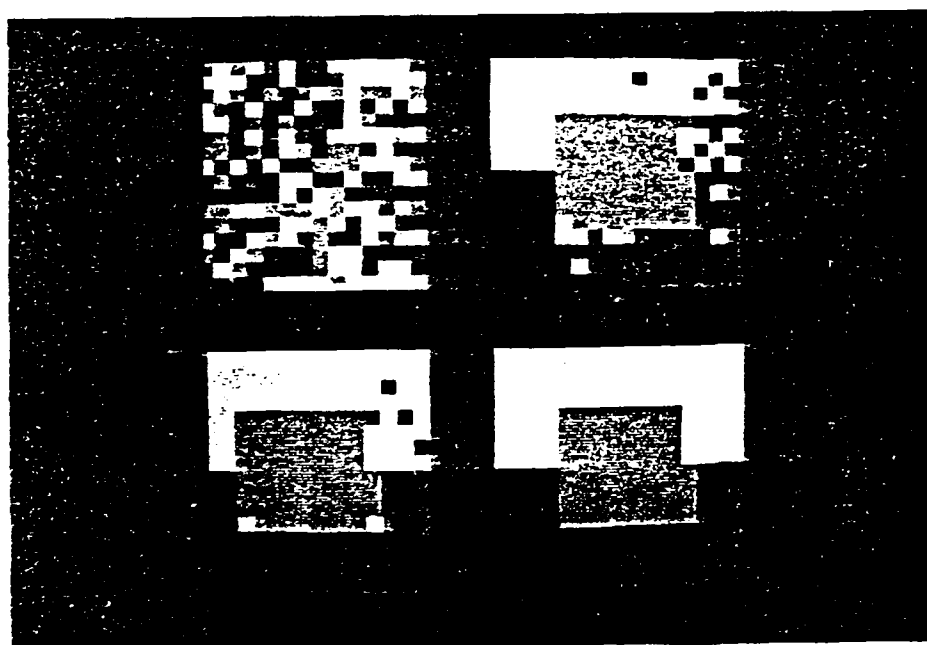


Figure 10(B)

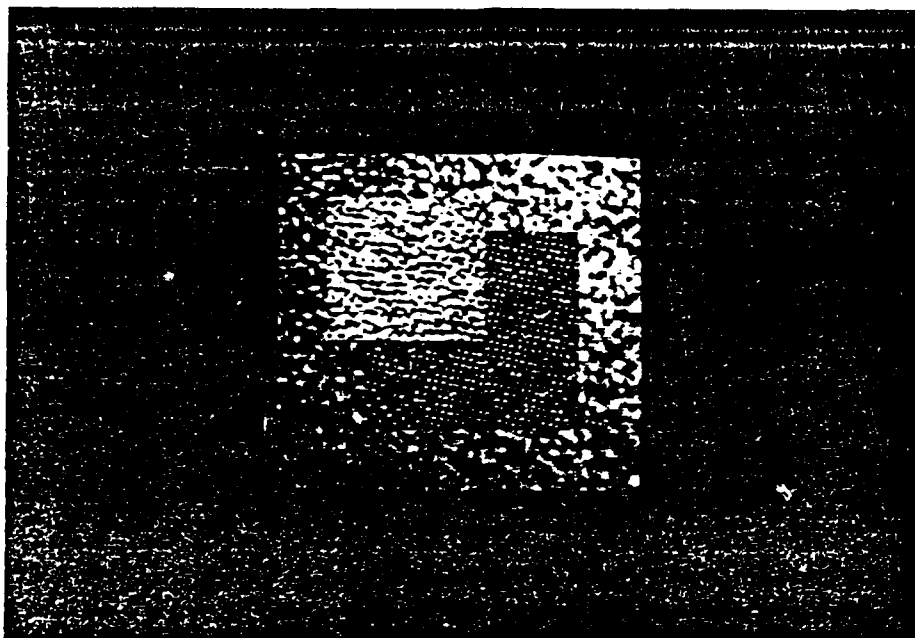


Figure 11(A)

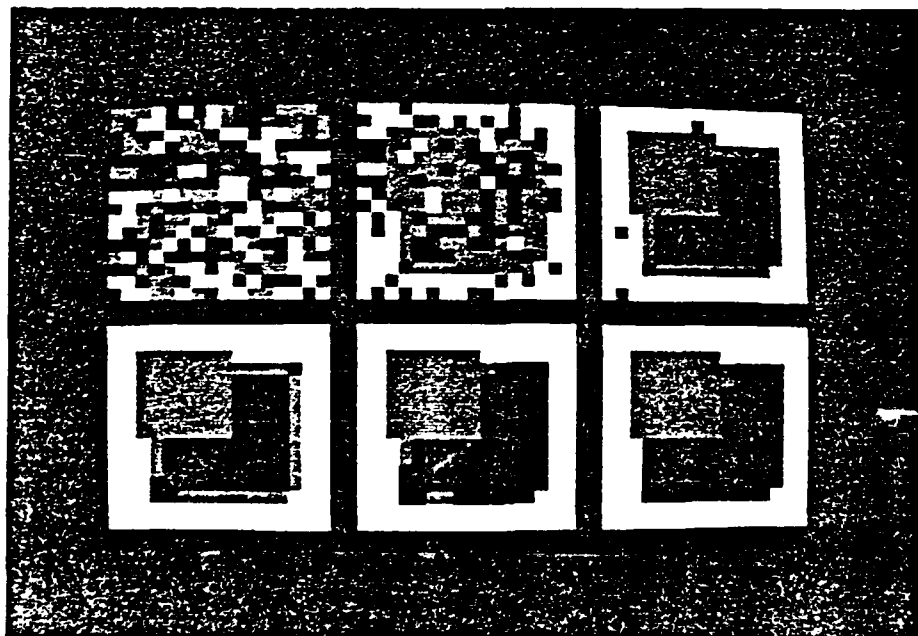


Figure 11(B)

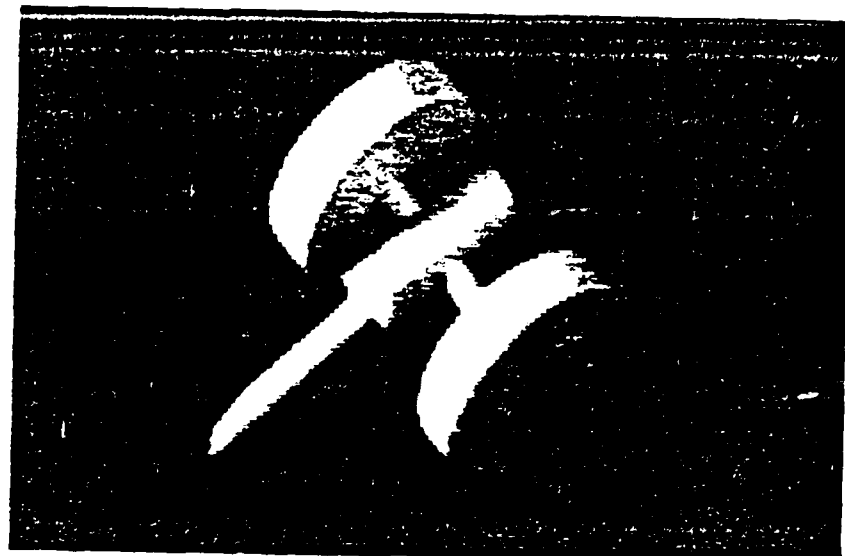


Figure 13(A)

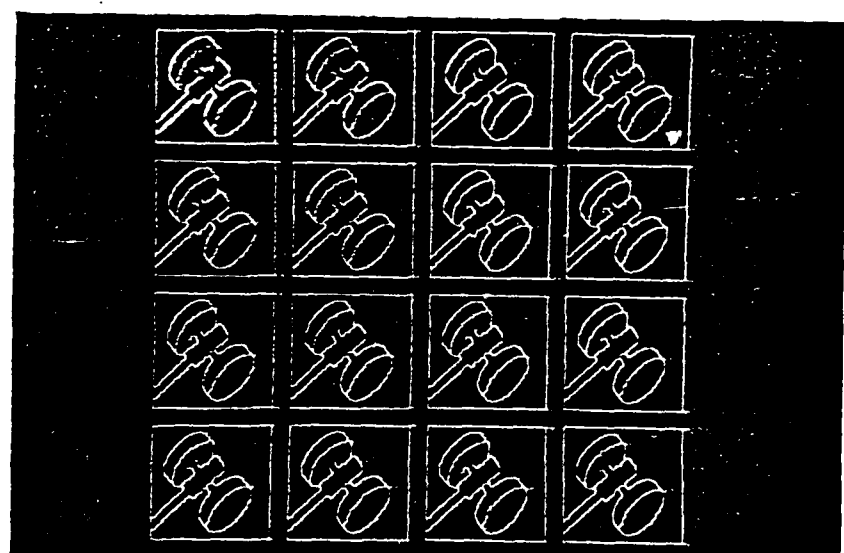


Figure 13(B)

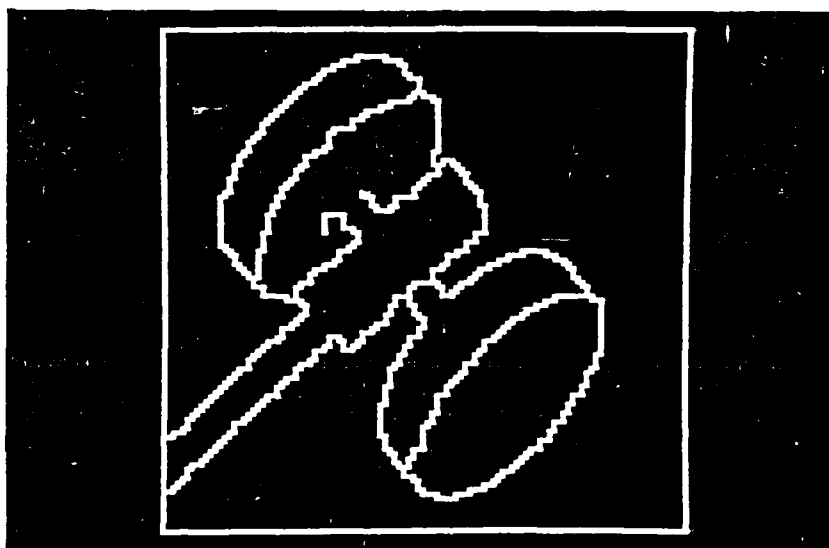


Figure 13(C)

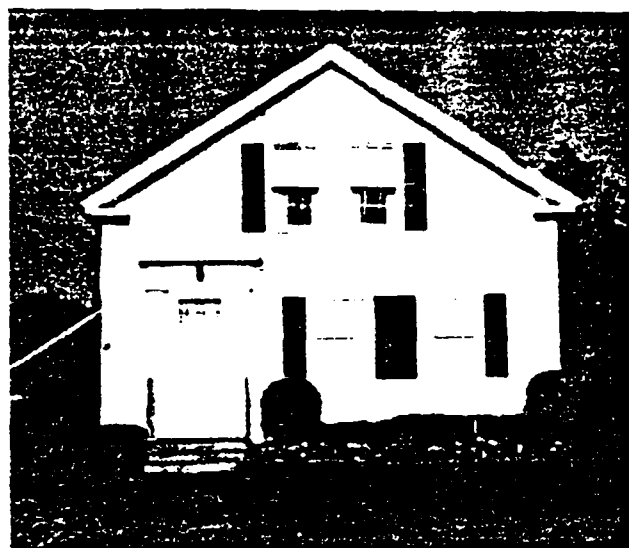


Figure 14(A)



Figure 14(B)

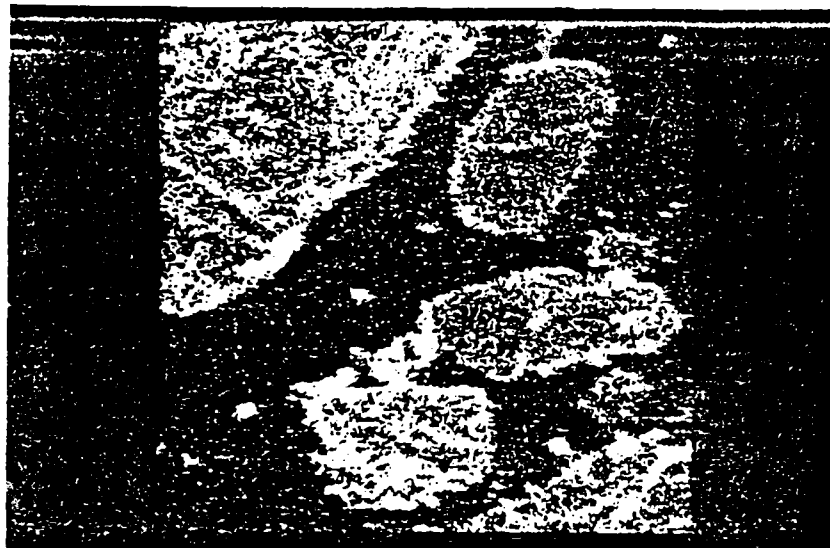


Figure 15(A)

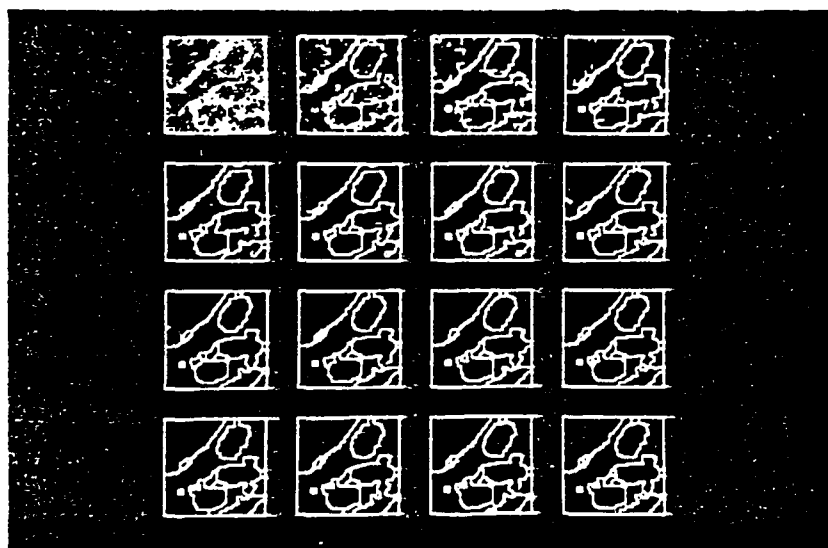


Figure 15(B)

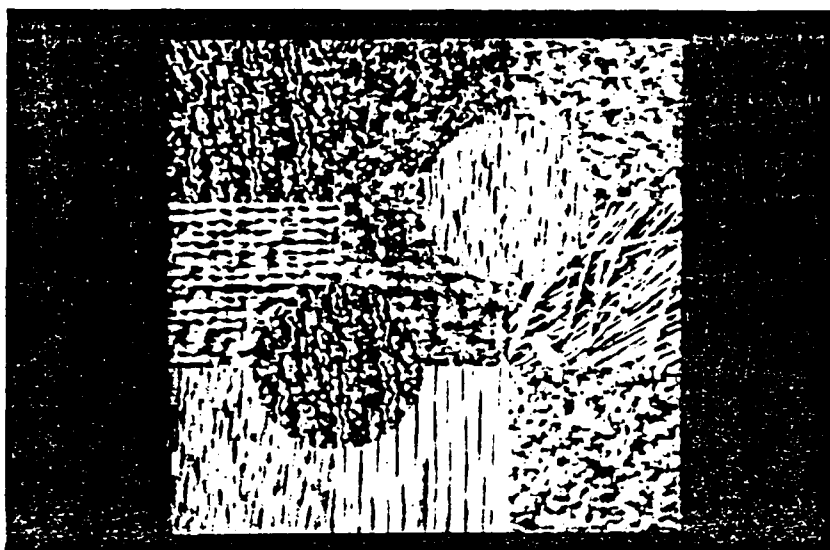


Figure 16(A)

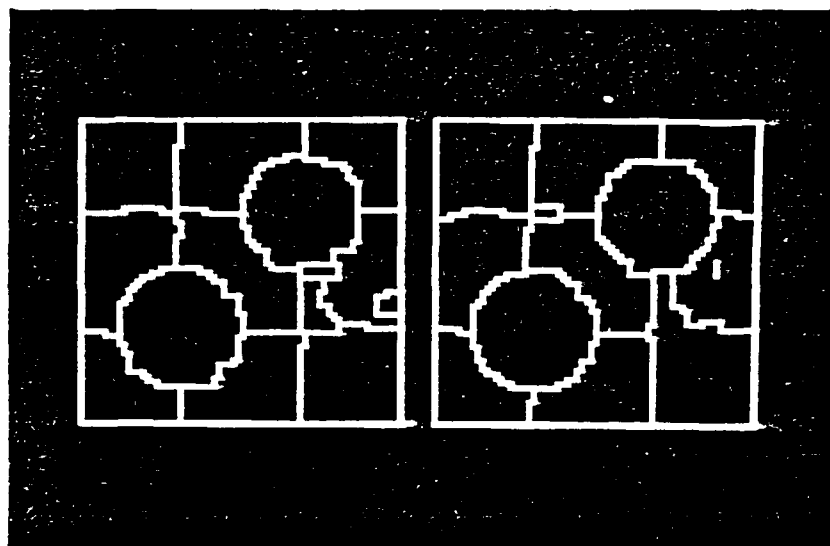


Figure 16(B)

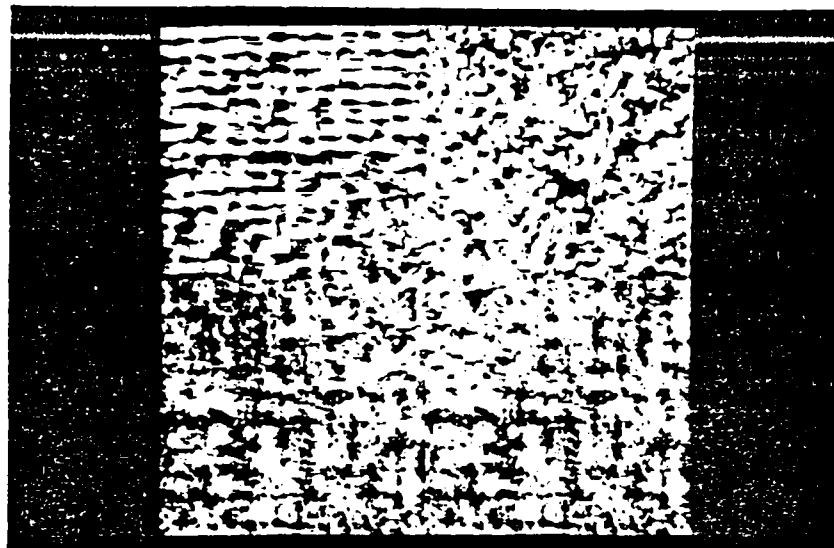


Figure 17(A)

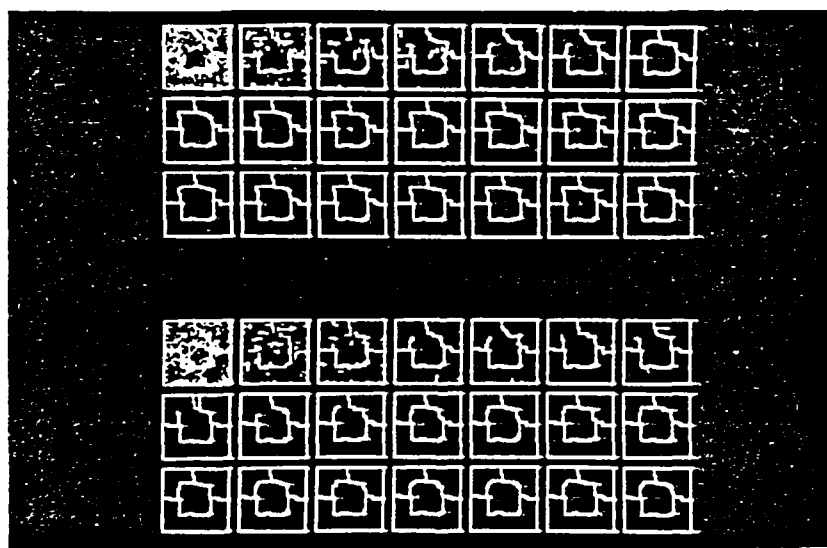


Figure 17(B)

MODELING VARIATION TO ENHANCE QUALITY IN MANUFACTURING

PRESENTED AT THE CONFERENCE ON
UNCERTAINTY IN ENGINEERING DESIGN

GAITHERSBURG, MARYLAND

MAY 10-11, 1988

BY HILARIO L. OH

CPC GROUP

GENERAL MOTORS CORPORATION

WARREN, MICHIGAN 48090

I appreciate this opportunity to show you an approach that we at CPC have developed to enhance quality in manufacturing through modeling.

To illustrate some key ideas in this approach, I will start with a simple example: the design of a wheelcover.

I will then go into greater details the mathematics of this approach; followed by yet another application on a more complicated example: the design of a door hanging process that consistently yields the same door closing effort.

I will then close with some comments.

OUTLINE

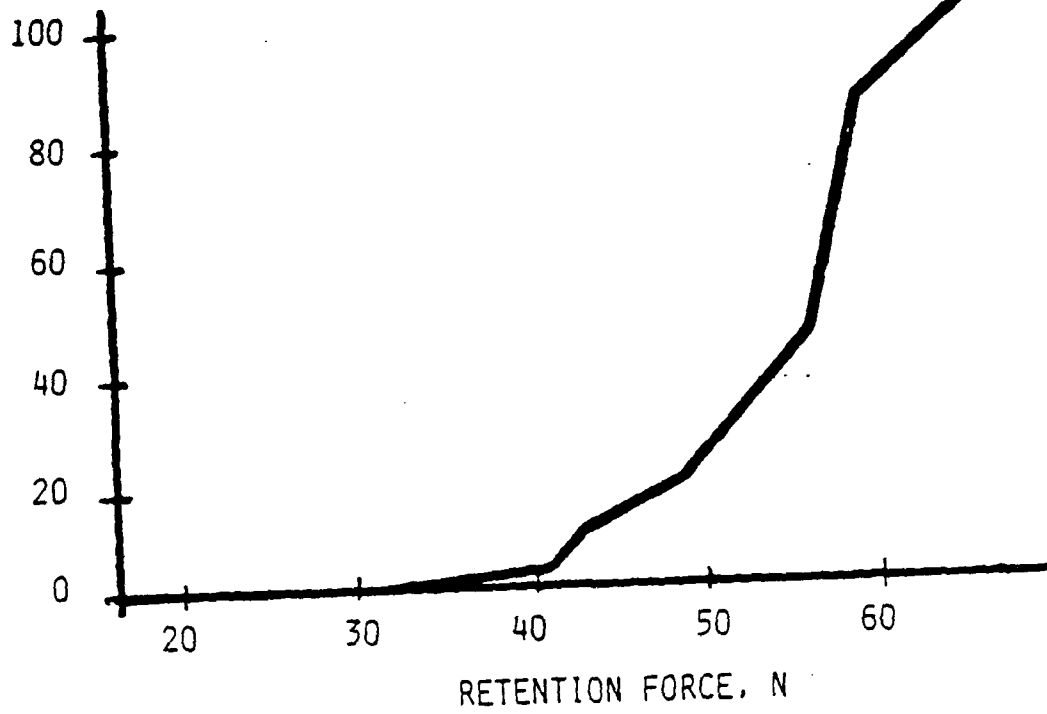
- ILLUSTRATE KEY IDEAS: DESIGNING A WHEELCOVER FOR CONSISTENCY
- MATHEMATICAL FRAMEWORK
- ANOTHER APPLICATION: DESIGNING DOOR HANGING PROCESS TO ACHIEVE CONSISTENT DOOR CLOSING EFFORT
- CONCLUSIONS

This is a typical wheelcover (show wheelcover). It is a simple and small part typical of the thousand parts that go in to make a quality car. In our business, the top priority is customer satisfaction. For this wheelcover, there are at least two features a customer has come to expect from it: ease of removal if you have to change the tire; and good retention on the wheel so you won't lose it when you hit a bump or turn a corner.

Depending on the customer, a male or a female, and on the tool used to remove the wheelcover, the retention force above which the cover becomes difficult to remove will vary from say 30 N, easily removed; to 60 N, completely unremovable and therefore 100% unsatisfactory.

WHEELCOVER REMOVAL

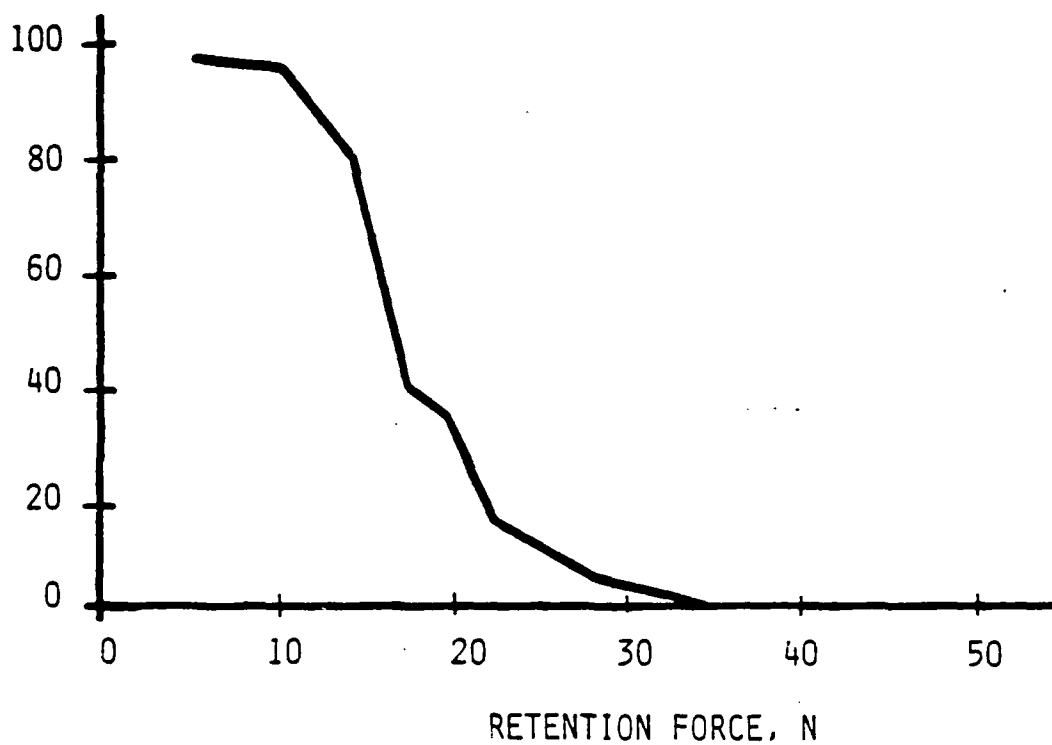
PERCENT OF
CUSTOMERS
DISSATISFIED



On the other hand, the retention force below which the cover will fall off also depends on the customer usage of the car. Cars on bumpy roads and with sharp turns require a higher retention force than cars on freeway driving. In other words, customer expectation on good retention will vary.

WHEELCOVER RETENTION

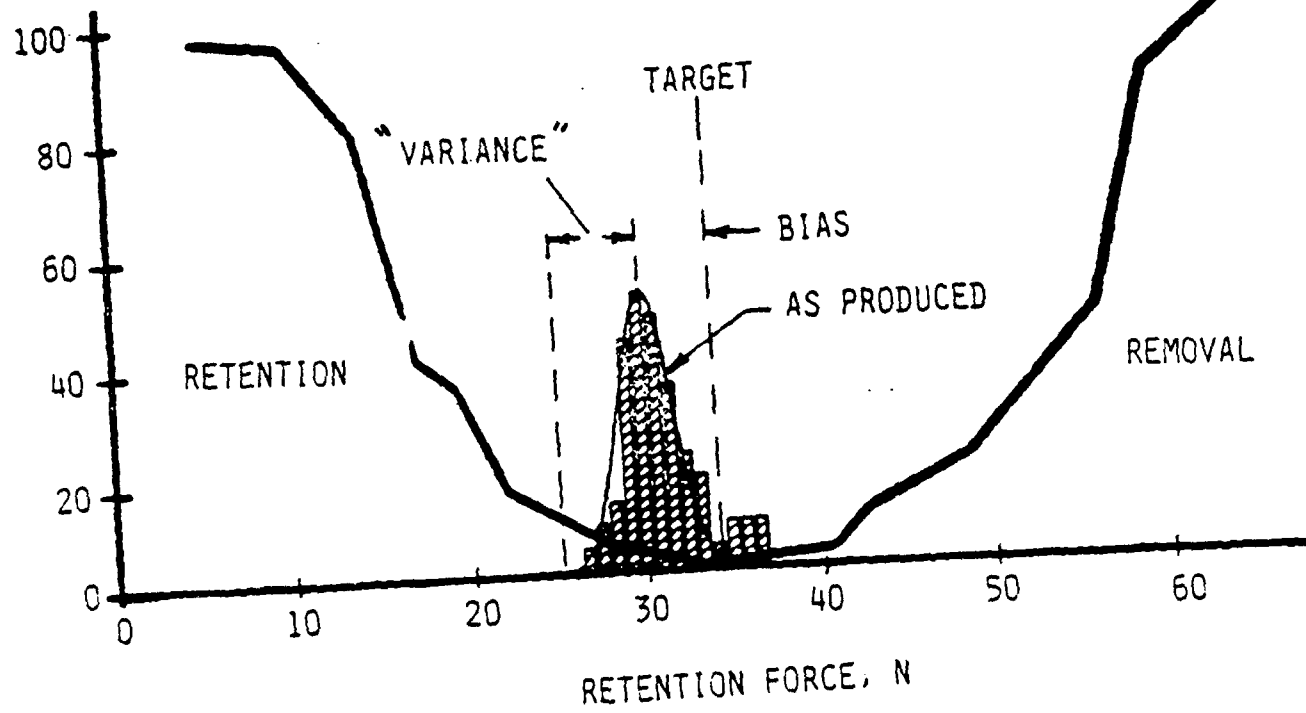
PERCENT OF
CUSTOMERS
DISSATISFIED



These two competing requirements combined into a target value of retention force deemed most satisfactory to the customer. Any departure from the target value will incur some degree of customer dissatisfaction and potential loss of market share.

COMPETING GOALS

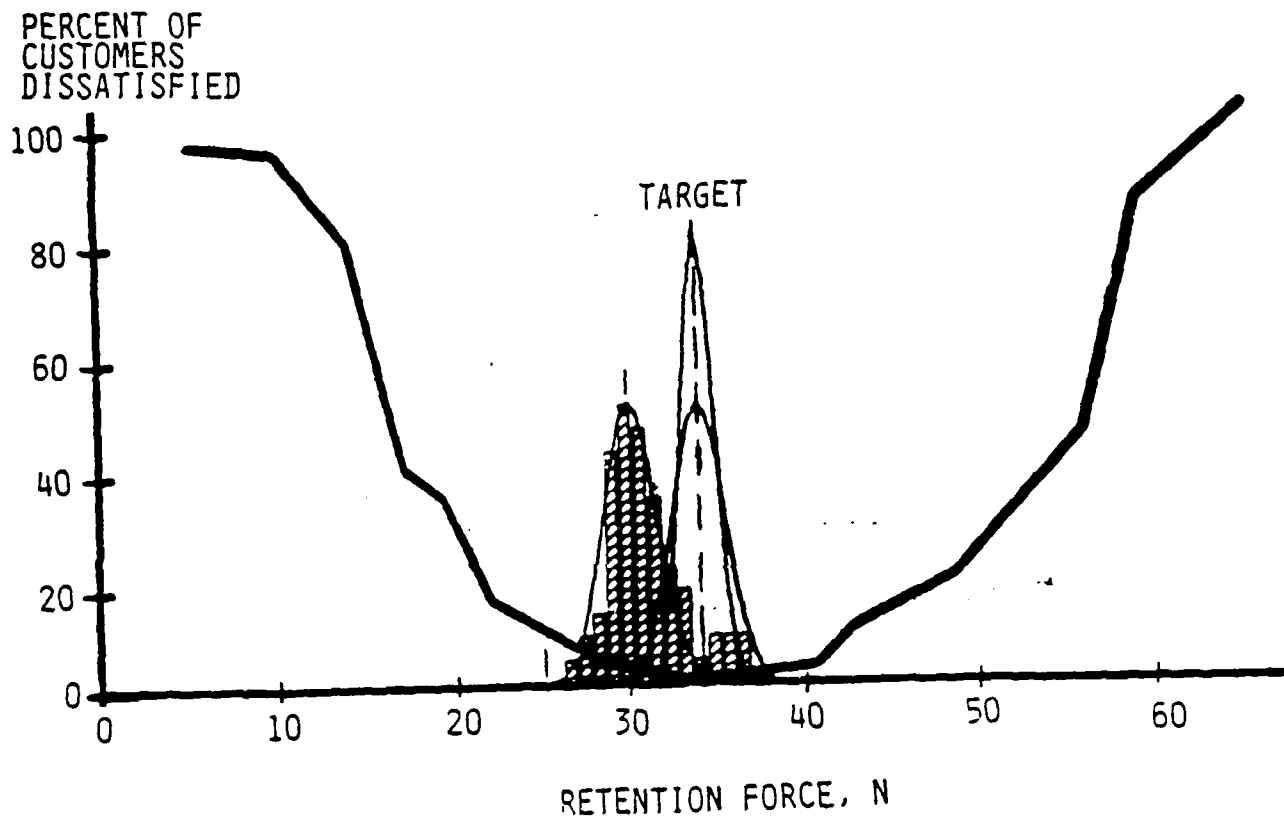
PERCENT OF CUSTOMERS DISSATISFIED



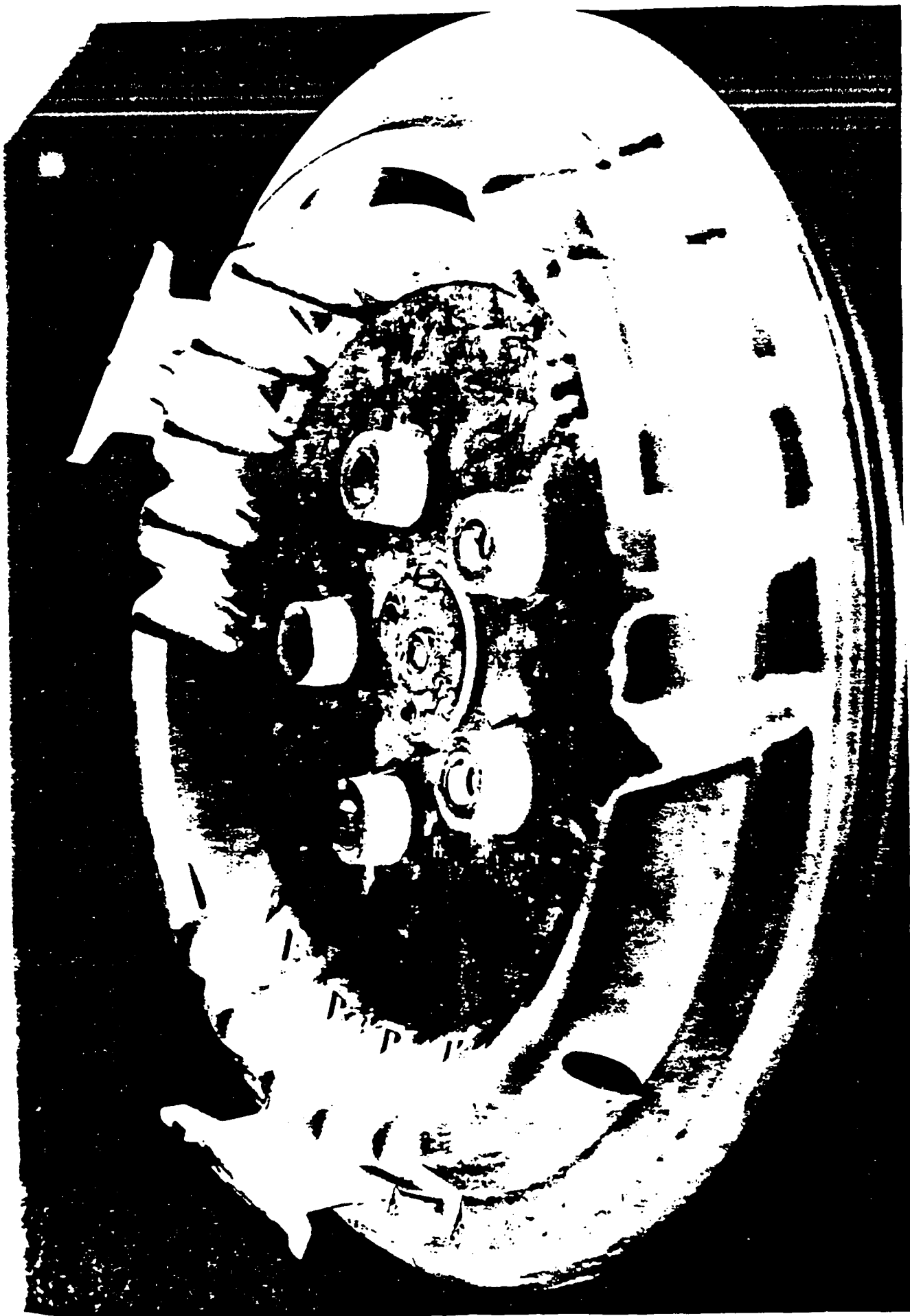
In our business of mass production, we will never be on target all the time. Most assuredly, we will produce cover with a range of retention forces whose mean is off target and a spread of values about the mean. The mean shift is called the bias; and the spread is called the variance. Our tasks are: (1) to get the mean on target -from an engineering viewpoint, that is not difficult-; and (2) to reduce the spread around the mean - that is difficult.

As a starting point, we must identify what are the factors or variables that affect the retention force.

OPERATIONAL GOAL

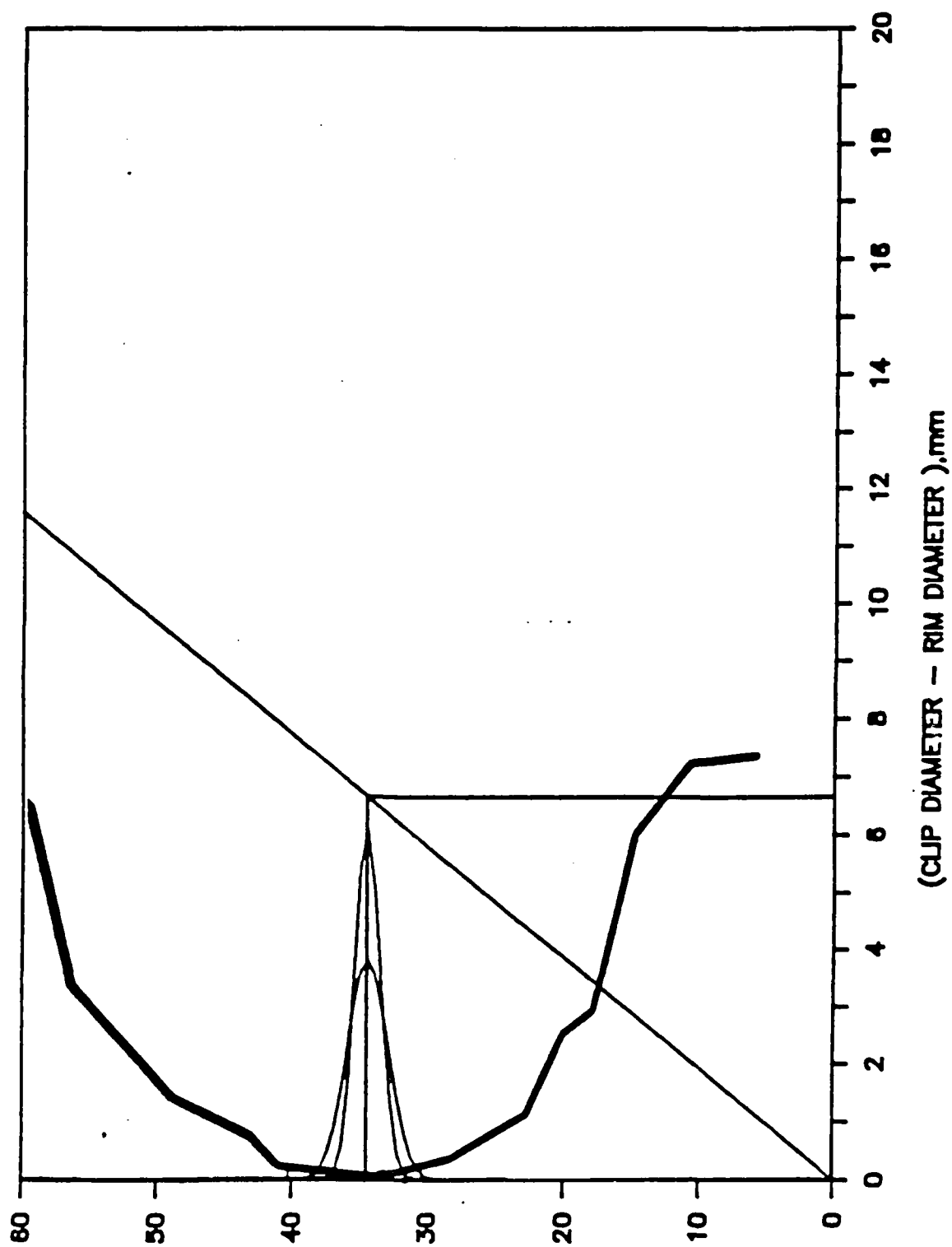


This slide shows back side of a typical wheelcover. The cover has three clips, each with two prongs, spaced around the circumference to form a circle. The diameter of this circle, which I call the clip diameter, is larger than the diameter of the rim on the wheel. So when you press the cover onto the wheel, the clip acts like a spring and clicks onto the rim.



With an understanding of the physics, we can now deploy customer expectation, retention force on target with minimum variation, in terms of engineering variables. There are two. The first one is the difference between the clip diameter and the rim diameter. Remember that the clip diameter is larger than the rim diameter. The larger the difference in diameters, the larger is the force developed. In fact, the relationship is linear. The second factor is the stiffness of the clip which is the slope of the line. As I mentioned earlier, it is easy to get the mean on target. For example, a difference in diameters of 6.65 mm and a clip stiffness of 5.2 N/mm will get us there.

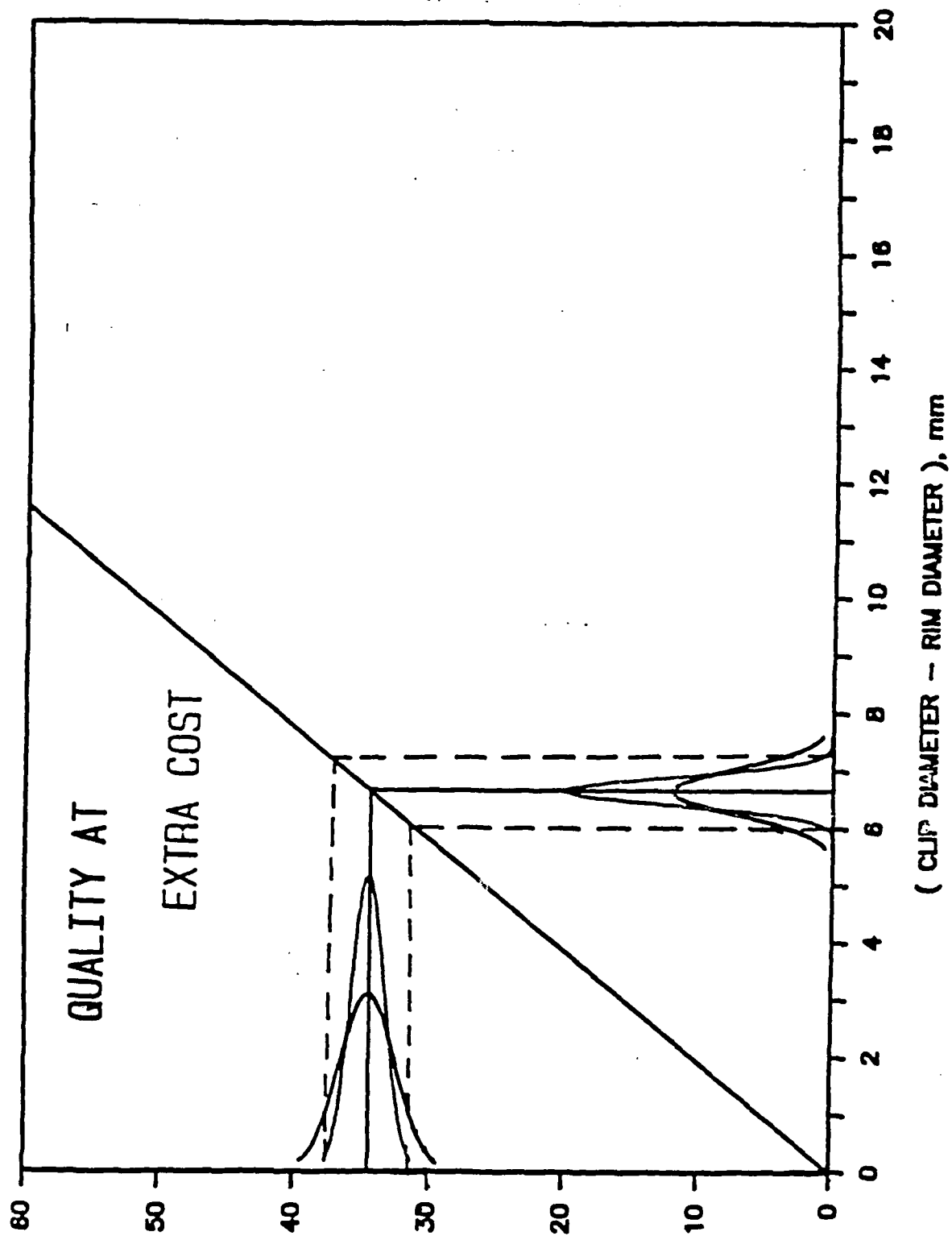
RETENTION FORCE, N



In our business of mass production however, there is always variation. We never get exactly 6.65 mm but a distribution of values instead. This distribution projects into a distribution of retention forces and results in less than satisfied customer. This is one point we all must realize -- in mass production, variation is a fact of life. It is the underlying cause of poor quality.

The usual practice is to control the variation by tightening the tolerance, say by sorting large covers to match with large rims and small covers with small rims. That of course, is expensive. So we achieve quality at extra cost, hoping to recover that cost through warranty cost reduction and improved customer satisfaction.

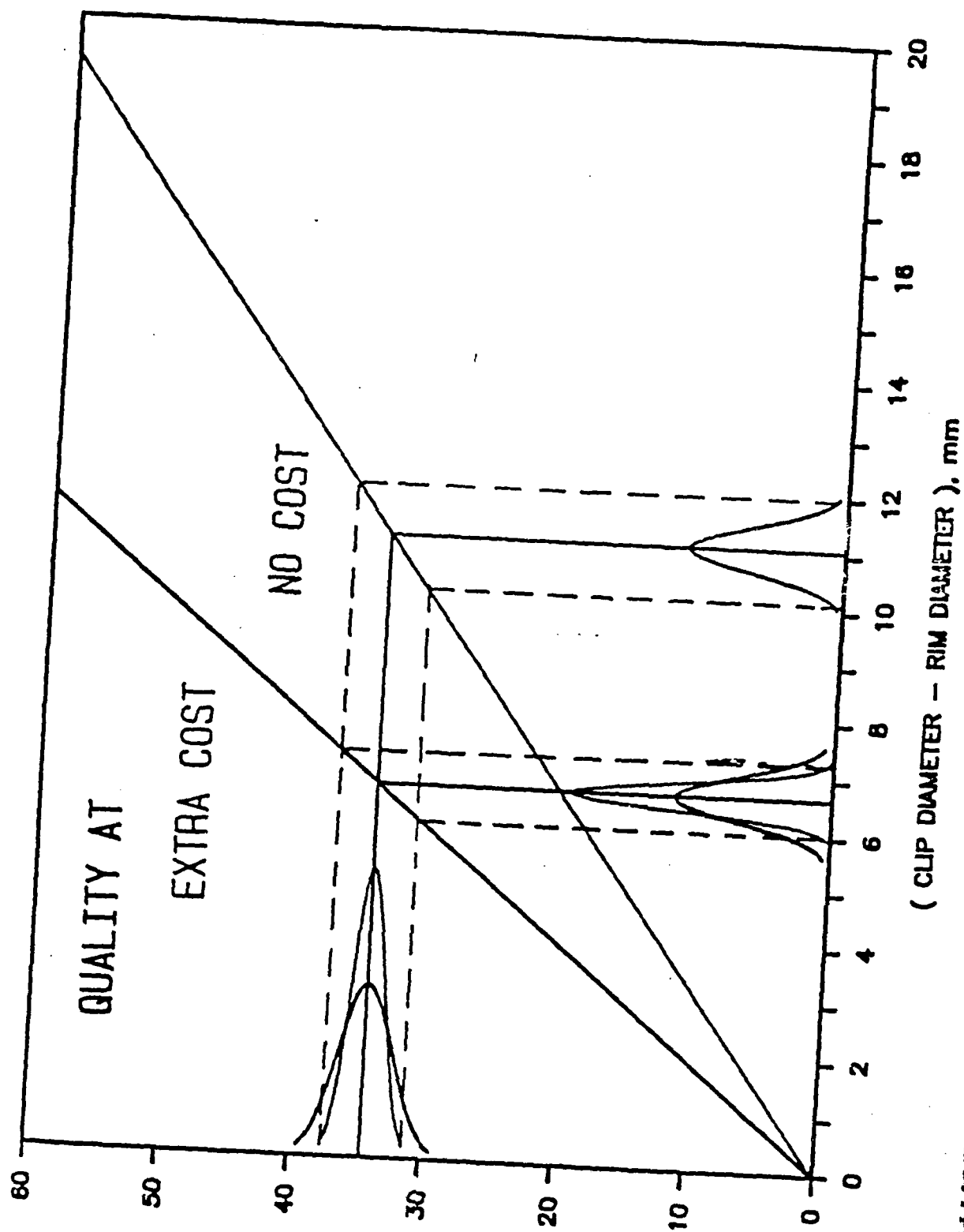
RETENTION FORCE, N



There is however, another approach. Instead of trying to sort the covers into large and small, we could choose less stiff clips spaced at a larger diameter. As you can see, with this choice, we can achieve the same quality with no cost because we do not have to tighten the tolerance. This is the Taguchi concept of insensitive design. It says: do not fight variation head on. Instead, make your design less sensitive to the variation.

At this stage, the warranty cost comes down and the customer is satisfied. But we, as engineers, should not be satisfied. Because ...

RETENTION FORCE, N

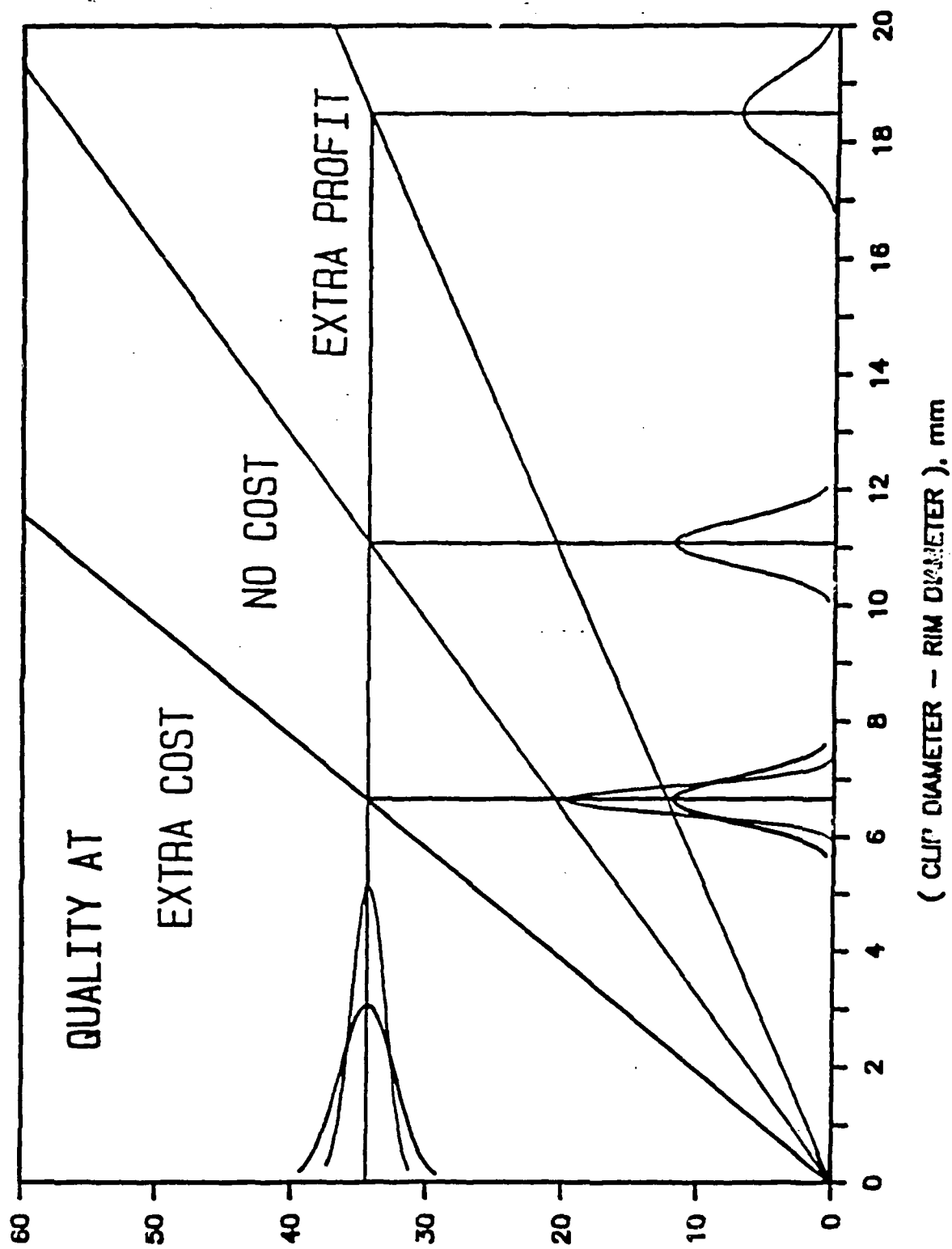


by choosing still weaker clips and a larger clip diameter, we can open up the tolerance and still arrive at the same quality. Now the cost really comes down. We can go with less expensive suppliers; our processes can be less precise; and the labor need not be skilled. In other words, we now achieve quality at extra profit; or "better for less". That -better for less- is the primary motivation behind the concept of insensitive design. Warranty cost reduction is only a secondary by product. It comes naturally when we do our design right.

Since better for less is our goal, our approach to process/product design must change. In this problem for example, the usual practice is for a design engineer to decide on the nominal clip diameter and stiffness to get the nominal retention force on target; and the manufacturing people would then decide what tolerances should go with them. Herein lies the crux of the problem: when the design engineer specifies the nominal clip diameter and stiffness, he already fixes the sensitivity of the wheelcover to the variation of mass production. The only mean left to the manufacturing people to reduce variation in retention force, if he wants to, is to tighten the tolerances of the clip diameter and stiffness, which as I said earlier is expensive. The more cost effective approach is the other way around. Go first to the manufacturing plant and negotiate for the most cost effective tolerance; and then come back and decide what the nominal clip diameter and stiffness should be to arrive at the quality we want.

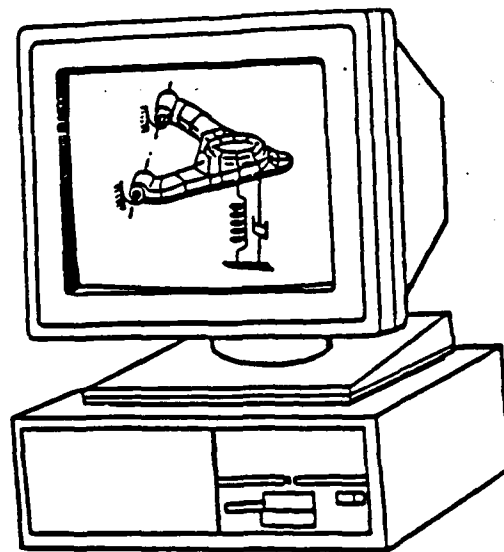
To summarize, what we have here is a miniature QFD, Quality Function Deployment, in which we have deployed customer expectation in terms of product/process variables. Now we can explore very early on in the design phase, the different product/process design alternatives that will produce quality product.

RETENTION FORCE, N



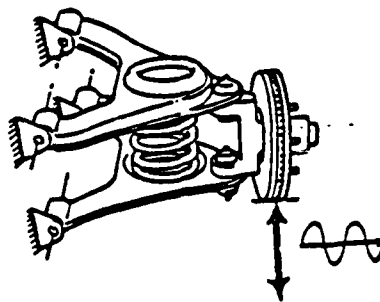
This sketch illustrates what we are trying to do. In on-line quality control, we try to inspect quality into the product through SPC and related statistical tools. I think that is too late and too expensive. In Taguchi method and related design of experiment methodology, we attempt to figure out early in the development phase what factors adversely affect quality and dissensitize the design against these factors. This is a big step forward. But I think that also is too expensive and too late. What I propose is to carry the activities further upfront in the design phase. Only then can we achieve the greatest impact.

Let me show you then the mathematical framework for carrying out this strategy.

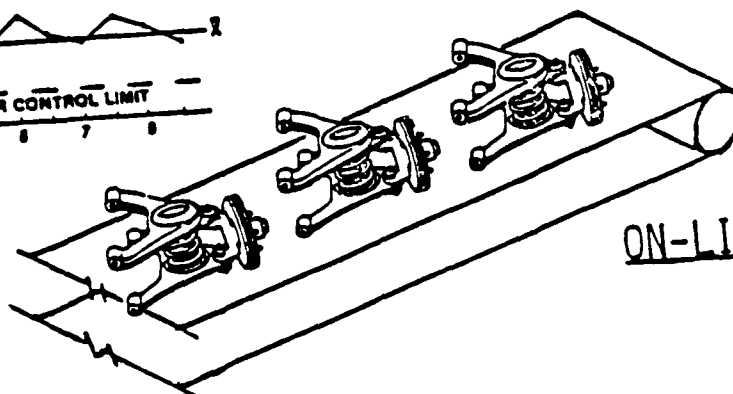
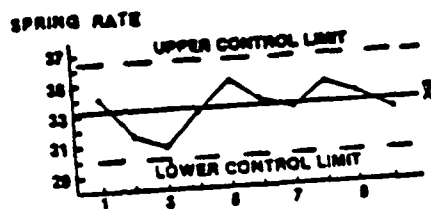


OFF-LINE QC
ANALYTICAL

Inner Array		Outer Array	EXPERIMENTS
1	2	3	
1	1	1	81
2	1	2	
3	1	3	
4	2	3	



OFF-LINE QC
EXPERIMENTAL



ON-LINE QC

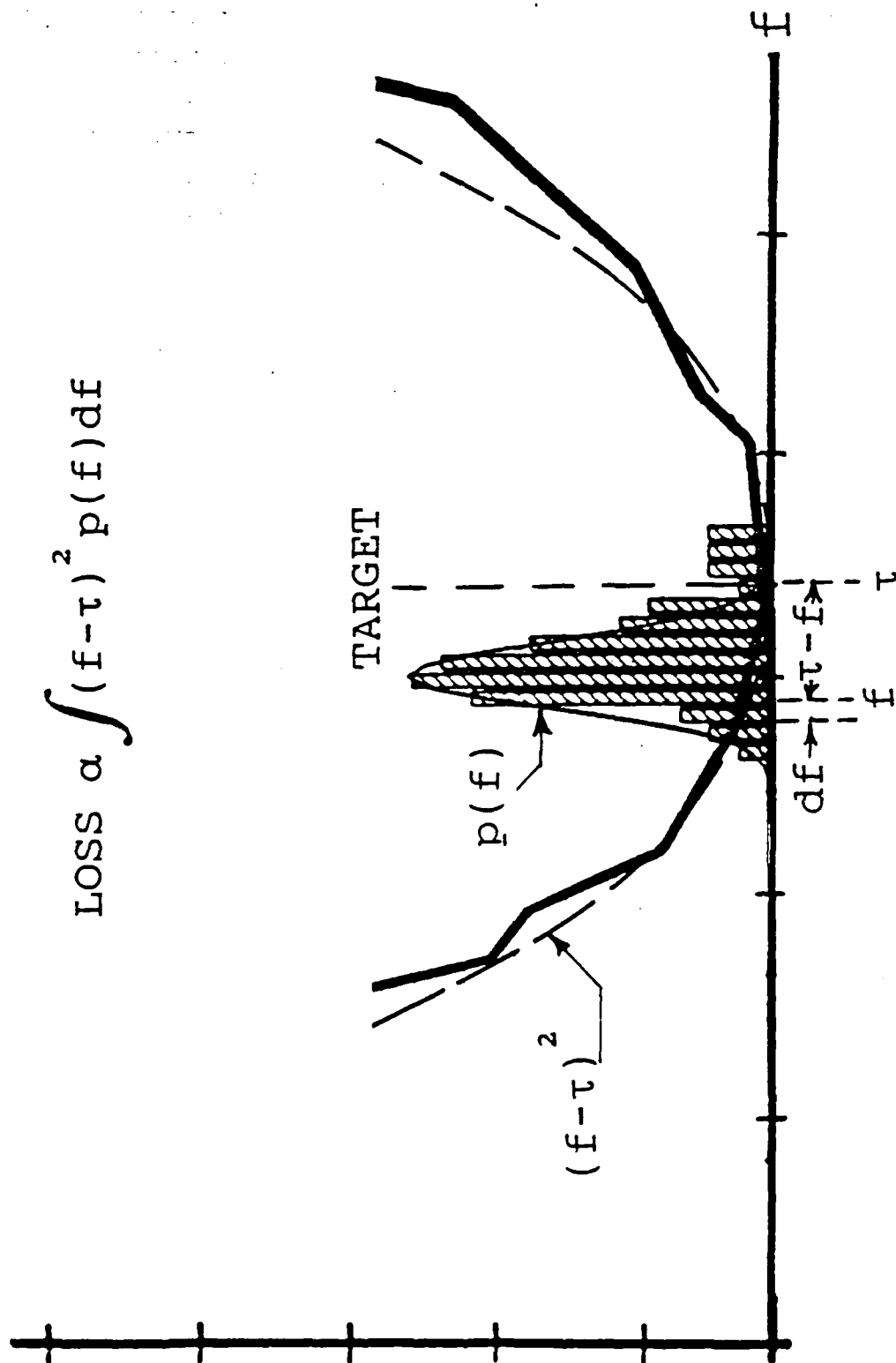
Let's go back to the wheelcover problem. In the neighborhood of the target τ , we may approximate customer dissatisfaction as a quadratic function of the retention force f , about the target. Assuming loss to be directly proportional to customer dissatisfaction, the fraction $p(f)$ of wheelcover population with retention force f that deviate from the target value entails a loss proportional to

$$(f-\tau)^2 p(f) df$$

Summing up this loss over the range of f , we have loss directly proportional to the mean squared error (MSE) of f .

$$\text{Loss} \propto \int (f-\tau)^2 p(f) df = \text{MSE}(f)$$

$$\text{LOSS} \propto \int (f - \tau)^2 p(f) df$$



The mean squared error (MSE) can further be decomposed into the square of the bias, a measure of how far the mean is off target, and variance, the spread of the response from its mean. Our aim is to get the mean on target and the variance, a minimum. The procedure of searching for design with this property, Taguchi called it parameter design. In traditional optimization, it is called equality constraint optimization. So what I am doing is just transferring technology. Whereas Taguchi method implements the concept in the development phase through design of experiment in the lab, I would implement the same concept in the design phase through optimization in the computer.

$$\text{MSE}(f) = E[(f-\tau)^2]$$

$$= E\{[f-E(f)]^2\} + [E(f)-\tau]^2$$

variance bias

The basic requirement is that we must know the relation between the response f and the factors x which control or affect f . In the case of the wheelcover, the response is the retention force, and the factors are the clip stiffness x_1 , the clip diameter x_2 and the rim diameter x_3 . The key is to expand the response function about the nominal values of the factors. You can then derive in closed form the bias and variance of the response in terms of the nominal values of the factors x . In these equations, g_i is the gradient of f with respect to the factor x_i , h_{ij} is the hessian of f with respect to x_i, x_j ; μ_i is the mean or nominal value of x_i ; and σ_{ij} is the variance-covariance matrix of x_i, x_j .

These equations relating bias and variance to factors x are then submitted to an optimizer to search for nominal values of x that ensure the response is on target and with minimum variance.

I would like to point out again the difference between the traditional design and the variation minimum design. Traditionally, the design engineer takes a deterministic approach and uses the bias relation to find the set of nominal values μ that ensures the response is on target. In so doing, he fixes the sensitivity $g_i(\mu)$ of the design to the variation of mass production. The manufacturing engineer, if he wants to improve the quality, has no other recourse but to tighten the tolerance, σ_{ij} which generally is expensive. By contrast in variation minimum design, one tries to find the set of nominal values that ensures not only the response is on target but also the variance is a minimum. And so quality is achieved at no cost because you don't have to tighten the tolerance. Indeed, it is possible to find a set of nominal values that even allows you to open up the tolerance. At that stage, we attain the 'better for less' situation.

$$f(x) = f(\mu) + \sum_{i=1}^n g_i(\mu)(x_i - \mu_i) \\ + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n h_{ij}(\mu)(x_i - \mu_i)(x_j - \mu_j) + \dots$$

$$\text{bias}(f) = f(\mu) - \tau + \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n h_{ij}(\mu) \sigma_{ij};$$

$$\text{var}(f) = E\{[f - E(f)]^2\} \\ = \sum_{j=1}^n \sum_{i=1}^n g_i(\mu) g_j(\mu) \sigma_{ij}$$

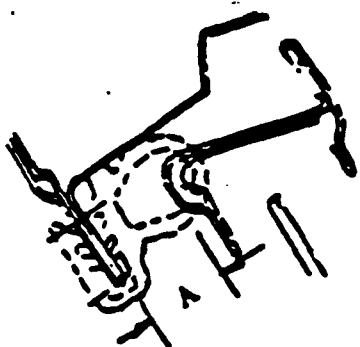
Let me demonstrate how this approach is implemented in another real life example: design a door hanging process that yields consistent door closing effort.

AN EXAMPLE

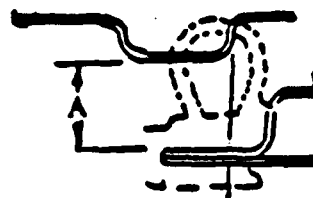
Design a door hanging process that yields consistent door closing effort.

Door closing effort in car is spent mostly in overcoming air resistance. Thus, much less effort is needed to close a door with the window down than with the window up.. We can reduce air resistance by closing the door slowly. However, the door needs to attain a certain velocity at closing for it to have the necessary amount of kinetic energy to compress the weatherstrip and effect a seal. The primary variable associated with door closing effort therefore, is the energy stored in the deformed weatherstrip. A reduction in stored energy (SE) means a reduced door velocity needed at closing which translates to a dramatic decrease in air resistance and door closing effort.

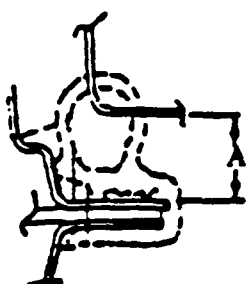
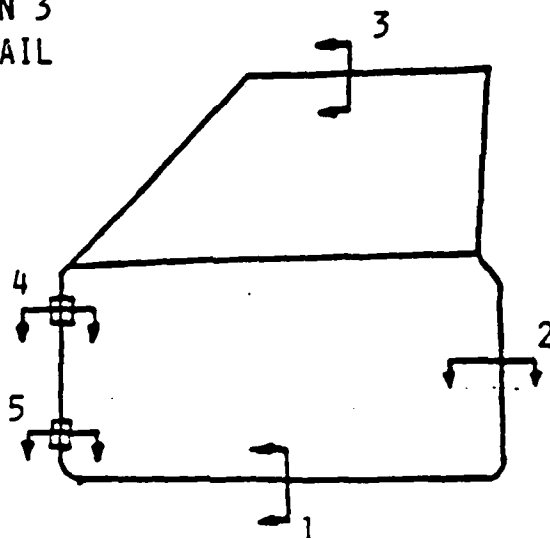
With the door closed, the weatherstrip is compressed between the body and the door around the door periphery as shown in this figure. The car to car variation in SE comes from the car to car variation in the diameter of the weatherstrip and in the gap between the door and the body. In turn, the variation in the gap comes from the variation in the build of both the body and the door and in the positioning of the door with respect to the body during hanging. For purpose of illustration, we consider only the variation due to hanging.



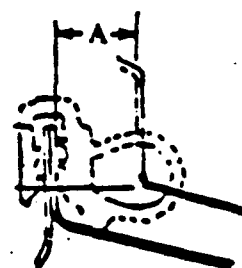
SECTION 3
ROOF RAIL



SECTION 2
LOCK PILLAR



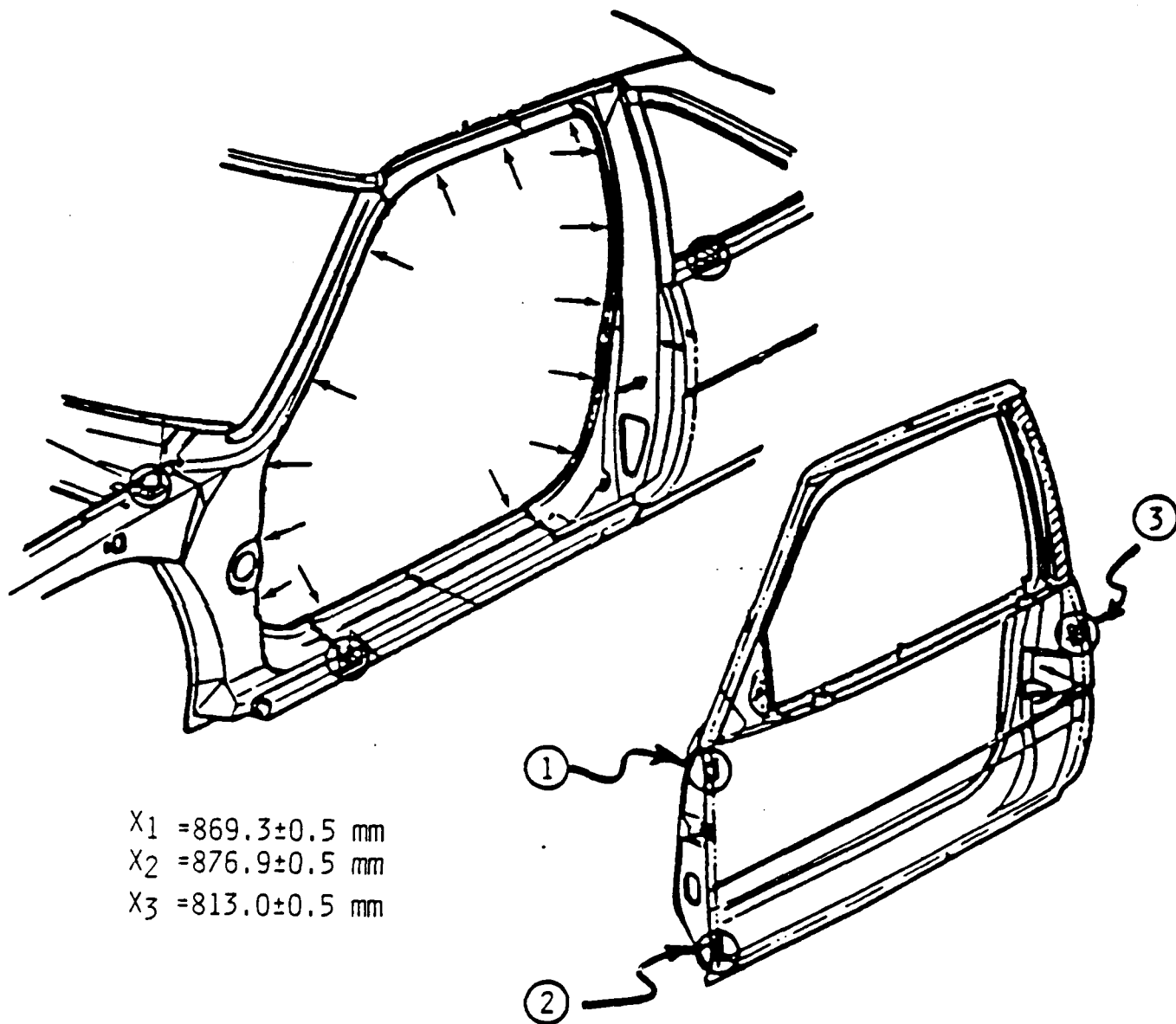
SECTION 4 & 5
HINGE PILLAR



SECTION 1
ROCKER

GAP 'A' AROUND THE DOOR PERIPHERY.

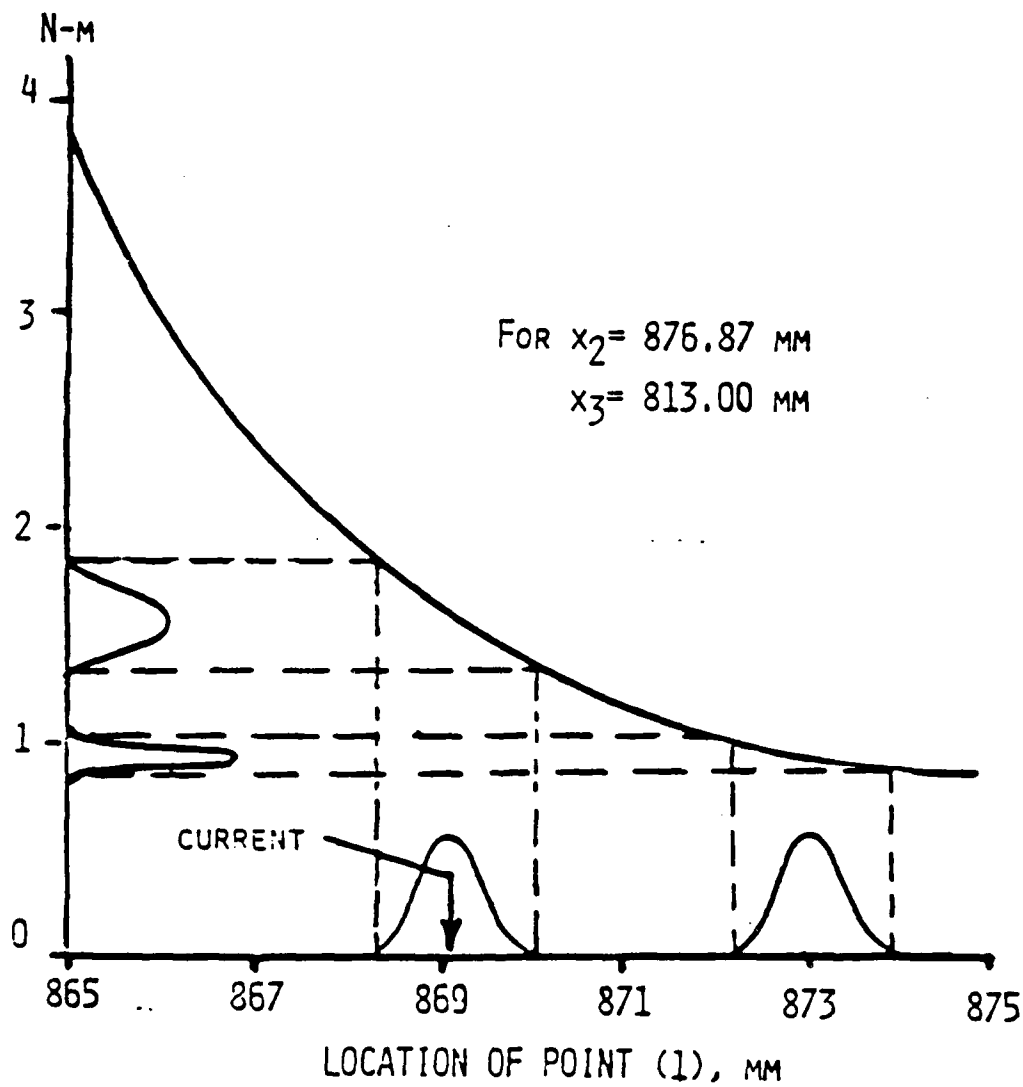
Take, for example, one car line currently being assembled. The door is positioned with respect to the door opening by locating three points (1,2,3) on the door at 869.26, 876.87 and 813.00 mm respectively, from the centerline of the car in the cross car direction as shown in this figure. These points correspond to the hinges and latch locations on the door. Once the door is positioned, hinges are screwed on to the door and the pillar. Suppose in positioning the door, the points (1,2,3) are in error by no more than 0.5 mm. How much deviation from the nominal value of the door closing effort do these errors produce? For the same tolerance of 0.5 mm, is there a trio of nominal cross car positions $\mu=[\mu_1, \mu_2, \mu_3]$ of the points (1,2,3) at which positions the door closing effort is the least sensitive to the errors in positioning?



POSITIONING OF THE DOOR TO THE OPENING

For example, this figure is a plot of the equation depicting how SE varies with x_1 , the position of the upper hinge. The lower hinge and the latch are fixed at their current nominal positions. The sensitivity of SE to hanging is now apparent. With x_1 at its current nominal position of 869.26 mm, a 1.0 mm deviation in x_1 produces a deviation in SE of about 0.25 N-m. By contrast, with x_1 at say 873.00 mm, the same amount of deviation in x_1 produces only a deviation in SE of about 0.1 N-m. It is this potential for desensitizing a design to the variation in manufacturing that we try to exploit: instead of trying to tighten the 1.0 mm deviation to a smaller value, we reconfigure the design, positioning x_1 at 873 mm, to render the design insensitive to variation.

STORED ENERGY
(CLOSING EFFORT)



STORED ENERGY VERSUS UPPER HINGE LOCATION.

To analyze this problem, we first derive from a physical consideration the response SE as a function of the factors $\mathbf{x}=[x_1, x_2, x_3]$, the cross car position of points (1,2,3). In a diametral compression, the weatherstrip exhibits a linear force-deflection relationship. Therefore, for a weatherstrip of diametral stiffness K, diameter D and length L, the function SE(\mathbf{x}) is:

$$SE(\mathbf{x}) = \frac{1}{2} \int K[D-G(s, \mathbf{x})]^2 ds, \quad \text{for } D > G;$$

where G is the gap which varies along the door periphery s.

We then discretized the weatherstrip into 15 segments L_i . The integral may now be approximated by summation, and the SE(\mathbf{x}) for a given \mathbf{x} may be computed:

$$SE(\mathbf{x}) = \frac{1}{2} K \sum_{i=1}^{15} [D-G_i(\mathbf{x})]^2 L_i;$$

where nominally, $K = 0.012 \text{ N/mm/mm}$, $D = 19\text{mm}$.

The above equation completely describes the relationship between the response SE and the factors \mathbf{x} .

STORED ENERGY (SE) IN WEATHERSTRIP

$$\begin{aligned} SE(x) &= \frac{1}{2} \int_0^L K[D-G(s,x)]^2 ds, \\ &= \frac{1}{2} K \sum_{i=1}^{15} [D-G_i(x)]^2 L_i \end{aligned}$$

We now exploit the full potential by considering simultaneously all three nominal positions $\mu=[\mu_1, \mu_2, \mu_3]$ of \mathbf{x} . Avoiding drastic departure from the current positions $\mu^c=[869.26, 876.87, 813.00]$, we search within a 2 mm neighborhood for a position μ^* at which the mean value of $SE(\mu^*)$ is the same as that of the current; and the variance of $SE(\mu^*)$ is minimized. Cast in the context of optimization, we submit the problem to a standard optimization routine and found the solution $\mu^* = [870.27, 874.87, 811.00]$. The result is shown in this figure. For the same 0.5 mm tolerance allowed in μ , the variance of SE at optimum position is only a third of that at current position.

The significance of this example is not so much in the results but in the fact that insensitive design concept can be integrated into CAD/CAM environment thus permits a widespread, early and upfront implementation of the concept.

$X_1 = 869.3 \text{ mm}$
 $X_2 = 876.9 \text{ mm}$
 $X_3 = 813.0 \text{ mm}$

$X_1 = 870.3 \text{ mm}$
 $X_2 = 874.9 \text{ mm}$
 $X_3 = 811.0 \text{ mm}$

CURRENT

OPTIMUM

1.0

1.4

1.8

STORED ENERGY, N-M

1.0

1.4

1.8

STORED ENERGY, N-M

HISTOGRAMS OF STORED ENERGY (CLOSING EFFORT) FOR
 DOOR HANGED AT CURRENT AND AT OPTIMUM POSITION.

To summarize, let me review these key points:

- that in mass production, variation is a fact of life;
- that we should not try to fight variation head on, but instead make our design insensitive to it;
- that in doing so, our primary motive is 'better for less';
- and finally, that a framework has been developed which relates customer expectation to product/process design; and allows a widespread, early and upfront implementation.

CONCLUSIONS

- VARIATION IS HERE FOREVER;
- DESENSITIZE THE DESIGN;
- PRIMARY MOTIVE IS "BETTER FOR LESS";
- FRAMEWORK DEVELOPED WHICH
 - RELATES CUSTOMER EXPECTATION
TO PRODUCT/PROCESS VARIABLES;
 - ALLOWS INTEGRATION INTO CAD/CAM & THUS
WIDESPREAD AND UPFRONT IMPLEMENTATION.

Ladies and gentlemen, in our business of mass production we look upon variation as an evil because it is the root cause of poor quality and unreliability. There is however, one thing good about variation--it is blind. It does not discriminate Ford from GM from Toyota. It affects everyone and exempts no one. For that reason we should view variation not as a problem, but as an opportunity to use it to our advantage and gain a competitive edge over our competitions.

**VIEW VARIATION
NOT AS A PROBLEM
BUT AS AN OPPORTUNITY**

BACKGROUND ARTICLE

VARIATION TOLERANT DESIGN

M. L. Oh
C-P-C Group
General Motors Corporation
Warren, Michigan

ABSTRACT

Taguchi philosophy of robust design is formulated in the context of optimization. In particular, the mean squared error of a design performance is derived explicitly in terms of the design parameters. This permits the strategic choice of the parameters that minimize the mean squared error of the performance early at the design phase. Thus, Taguchi philosophy will achieve an even greater impact on quality improvement as its implementation is shifted from the usual domain of experimental development to the early phase of analytical design.

A real life problem is used to illustrate the implementation of the formulation: design a door hanging process in car assembly that yields consistent door closing effort.

INTRODUCTION

Quality improvement activities achieve their maximum impact when they are carried out up front in the product realization process. Thus, the Taguchi philosophy of shifting quality control from assembly line inspection to pre-production experimentation is a significant step in this direction. Another contribution of Taguchi is the philosophy of robust design: do not try to control the sources of variation affecting the design; make the design insensitive to these sources instead.

To implement the Taguchi philosophy, one usually employs a planned experimental program to acquire knowledge about: (1) the relationship between the design response and the factors affecting the response; and (2) the variability in the design response caused by the variability in the factors. Based on the knowledge (1) acquired, one then sets the factors to some nominal values such that the average value of the design response equals a target value while the variance of the design response in (2) is at a minimum. In this way, one has achieved a variation minimum or variation tolerant design. A variation tolerant design may not achieve optimal performance level; but it will have performance

consistency. It is the latter that is to be emphasized to achieve robust design.¹

The strategy as described above is well known in the field of optimization and is called equality constraint minimization [1].² If the relationship between the response and the factors is known, then as pointed out in [2,3], variation tolerant design can be accomplished by implementing the Taguchi philosophy solely through numerical optimization. No experimentation is needed. This is the motivation of this paper: to show that traditional optimization techniques can be applied in the design phase for the design of performance consistency, as opposed to Taguchi method which tries to achieve performance consistency in the development phase through the design of experiments.

There are several compelling reasons to implement the Taguchi philosophy of robust design through numerical optimization. First, greater impact can be achieved since we are taking another step further up front; i.e., from experimental development to analytical design. Second, there already exist in the engineering sciences, a vast amount of knowledge relating design response to design factors. These relationships are either known or can be derived through a simple application of the physical laws or a sophisticated modelling such as finite element modelling. By tapping into this existing knowledge and evoking Taguchi strategy through optimization, more design alternatives can be explored in a shorter time; and much variation tolerant design can be achieved with little experimentation. Finally, with more and more products and processes now being designed with computer-aided engineering (CAE), using optimization to implement the Taguchi philosophy permits the integration of the Taguchi philosophy into CAE and allows the full realization of computer-aided robust design. In the next section, we devise a method which permits the implementation of the Taguchi philosophy of robust design in the context of optimization. In the last section, we use a real life example to illustrate the implementation: design a door hanging process in car assembly that yields consistent door closing effort.

FORMULATION

Let f denote the value of the design response of interest and τ , its target. As mentioned earlier, the dependency of f on the design factors $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ ³ is either known to us or can be derived. Because of the variability in \mathbf{x} , f would exhibit a random deviation from τ . Let the mean squared error (MSE) be a measure of this deviation. Then,

$$\begin{aligned} \text{MSE}(f) &= E[(f - \tau)^2] \\ &= E[(f - E(f))^2] + [E(f) - \tau]^2. \end{aligned} \quad (1)$$

¹Author acknowledges these terse statements by one discussor.

²Numbers in bracket refer to references at the end of the text.

³Bold symbols are vector quantities unless otherwise specified.

The symbol E denotes expectation, or averaging over x :

$$E(\cdot) = \int (\cdot) p(x) dx ;$$

where $p(x)$ is the joint probability density function of the design factors x . On the right hand side of Eq (1), the first term is the variance of f . This is a measure of the variation of f from its mean. The second term is the bias squared. This is a measure of the deviation of the mean value of f from the target. It is a fixed error affecting every design. Our aim is to minimize $MSE(f)$ by a choice of the mean values $\mu = (\mu_1, \mu_2, \dots, \mu_n)$ of x that:

$$\text{ensures zero bias} \quad E(f) - \tau = 0;$$

$$\text{and minimizes variance} \quad E\{[f - E(f)]^2\}.$$

Taguchi called the above strategy, parameter design. In optimization, it is called equality constraint minimization. If $p(x)$ is known, e.g., normal distribution with known standard deviations but with the means as parameters yet to be determined, the above strategy could be carried out numerically. The computation, however, is simplified considerably and no knowledge of $p(x)$ is required if we expand $f(x)$ about μ in Taylor series and neglect terms of order three or higher:

$$f(x) = f(\mu) + \sum_{i=1}^n g_i(\mu)(x_i - \mu_i) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n h_{ij}(\mu)(x_i - \mu_i)(x_j - \mu_j) \quad (2)$$

where $g_i(\mu)$, $h_{ij}(\mu)$ are the partial derivatives $\partial f / \partial x_i$, $\partial^2 f / \partial x_i \partial x_j$ evaluated at $\mu = (\mu_1, \mu_2, \dots, \mu_n)$. Since $E(x_i) = \mu_i$, the mean value of f is

$$E(f) = f(\mu) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n h_{ij}(\mu) \sigma_{ij} \quad (3)$$

and the bias and variance of f may be computed from Eqs (2,3) as follows:

$$\text{bias}(f) = E(f) - \tau = f(\mu) - \tau + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n h_{ij}(\mu) \sigma_{ij} \quad (4)$$

$$\begin{aligned} \text{var}(f) &= E\{[f - E(f)]^2\} = E\left\{\left[\sum_{i=1}^n g_i(\mu)(x_i - \mu_i)\right]^2\right\} \\ &= \sum_{i=1}^n \sum_{j=1}^n g_i(\mu) g_j(\mu) \sigma_{ij} \end{aligned} \quad (5)$$

In above equations, σ_{ij} denotes the variance-covariance of (x_i, x_j) given by

$$\begin{aligned} \sigma_{ij} &= E[(x_i - \mu_i)(x_j - \mu_j)] \\ &= \text{var}(x_i), \quad \text{for } i = j; \\ &= \text{cov}(x_i, x_j), \quad \text{for } i \neq j. \end{aligned}$$

Several philosophical points about variation tolerant design are evident from Eqs (4,5). In product or process design, the usual practice is to choose a set of μ such that the response f is on target. This is usually done based on deterministic calculations. Once the choice is made, the gradient $g_i(\mu)$ in Eq (5) is

fixed. This fixes the variance of f . If the variance of f is unacceptable, the design engineer would then attempt to reduce it by tightening σ_{ij} , the variance-covariances of the design factors in Eq (5). This approach generally requires more precise processing and skillful labor. It is a more expensive approach. Thus, quality is achieved at extra cost. Hopefully, this extra cost will be recovered in the form of reduced warranty cost and improved customer satisfaction.

By contrast, variation tolerant design attempts to find a set of μ not only to ensure that the response is on target but also to guarantee that the gradient $g_j(\mu)$ in Eq (5) is a small value. This would reduce the variability in the response f at no extra cost since there is no attempt to tighten σ_{ij} . Therefore, quality is achieved at no additional cost. Indeed, there exists the possibility of finding a set of μ which reduces $g_j(\mu)$ to such a low level that σ_{ij} may be opened up without increasing the variability in the response f . In this case, manufacturing becomes less costly and quality is achieved at extra profit. In short, the primary motivation in variation tolerant design is profit. Warranty cost reduction and improved customer satisfaction are secondary by-products.

A related point to the above discussion is the following. While the usual design practice is to first set nominal values of the design factors based on deterministic calculations and then let the design engineer and the production people negotiate on what the tolerances of these factors should be, the more profitable approach is to first set the tolerances of these factors at values that minimize the cost and then implement the variation tolerant design to choose the nominal values that ensure design response on target at minimum variability. As a corollary, the design engineer must know σ_{ij} , the capability of manufacturing before he can carry out a variation tolerant design.

AN EXAMPLE

Door closing effort in car is spent mostly in overcoming air resistance. Thus, much less effort is needed to close a door with the window down than with the window up. We can reduce air resistance by closing the door slowly. However, the door needs to attain a certain velocity at closing for it to have the necessary amount of kinetic energy to compress the weatherstrip and effect a seal. The primary variable associated with door closing effort, therefore, is the energy stored in the deformed weatherstrip. A reduction in stored energy (SE) means a reduced door velocity needed at closing which translates to a dramatic decrease in air resistance and door closing effort.

With the door closed, the weatherstrip is compressed between the body and the door around the door periphery, Figure 1. The car to car variation in SE comes from the car to car variation in the diameter of the weatherstrip and in the gap between the door and the body. In turn, the variation in the gap comes from the variation in the build of both the body and the door and in the

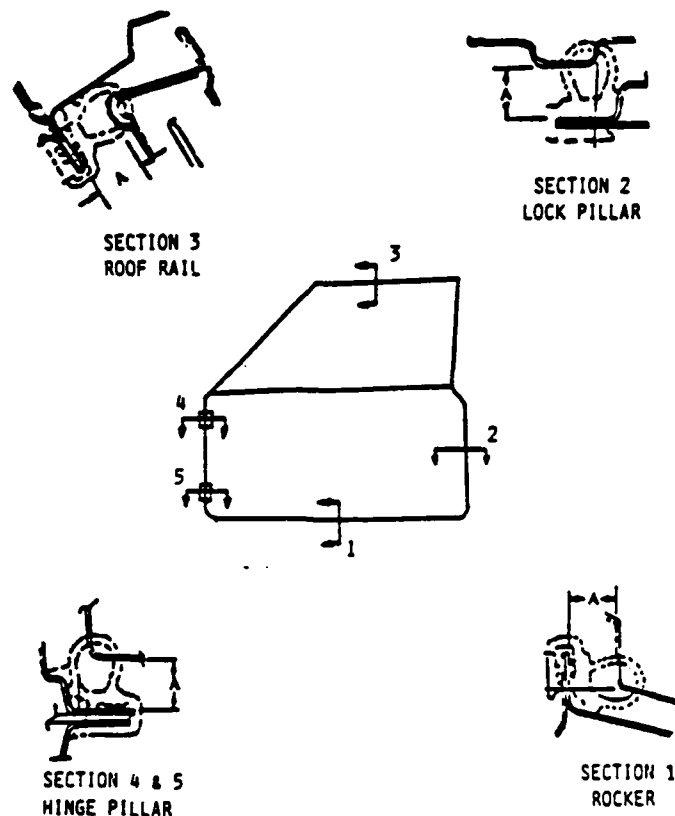


FIGURE 1. GAP 'A' AROUND THE DOOR PERIPHERY.

positioning of the door with respect to the body during hanging. For purpose of illustration, we consider only the variation due to hanging.

Take, for example, one car line currently being assembled. The door is positioned with respect to the door opening by locating three points (1,2,3) on the door at 869.26, 876.87 and 813.00 mm, respectively, from the centerline of the car in the cross car direction, Figure 2. These points correspond to the hinges and latch locations on the door. Once the door is positioned, hinges are screwed on to the door and the pillar. Suppose in positioning the door, the points (1,2,3) are in error by no more than 0.5 mm. How much deviation from the nominal value of the door closing effort do these errors produce? For the same tolerance of 0.5 mm, is there a trio of nominal cross car positions $u=[u_1, u_2, u_3]$ of the points (1,2,3) at which positions the door closing effort is the least sensitive to the errors in positioning?

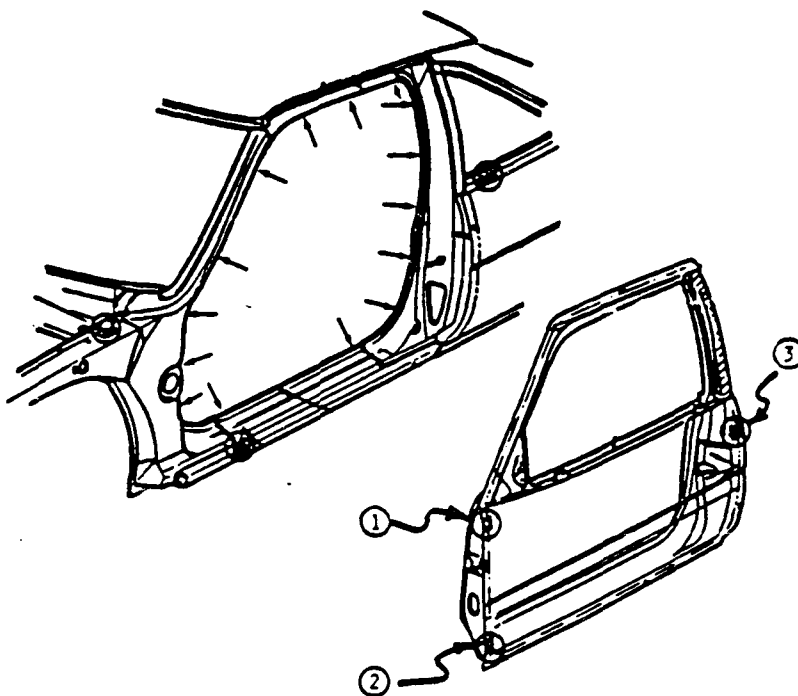


FIGURE 2. POSITIONING OF THE DOOR TO THE OPENING

To analyze this problem, we first derive from a physical consideration the response SE as a function of the factors $x = [x_1, x_2, x_3]$, the cross car position of points (1,2,3). In a diametral compression, the weatherstrip exhibits a linear force-deflection relationship. Therefore, for a weatherstrip of diametral stiffness K , diameter D and length L , the function $SE(x)$ is:

$$SE(x) = \frac{1}{2} \int_0^L K[D - G(s, x)]^2 ds, \quad \text{for } D > G; \quad (6)$$

where G is the gap which varies along the door periphery s . Since only small rotations are involved in positioning the door, the gap would be linearly proportional to x . Therefore upon integration, the $SE(x)$ will be a quadratic function of x :

$$SE(x) = a_0 + \sum_{i=1}^3 b_i x_i + \sum_{j=1}^3 \sum_{i=1}^3 c_{ij} x_i x_j. \quad (7)$$

The coefficients a_0 , b_i , and c_{ij} are estimated as follows.

Cartesian coordinates of two sets of points are digitized from the blueprints. One set of 15 points is on the body spaced around the door opening at locations indicated by the arrows, Figure 2. The other set of equal number of points is on the door directly opposite to the 15 points on the body. The set on the body is held fixed while the set on the door moves with the locating points (1,2,3) as a rigid body as the locating points are positioned to an x value. The coordinates of the i th point on the door as moved can be computed from its initial coordinates and the x values. Therefore, the distance between the i th point on the door and the opposing point on the body can be computed. This distance represents the gap G_i at that location of the door periphery. Approximating the integral in Eq (6) by summation, the $SE(x)$ for a given x may be computed:

$$SE(x) = \frac{1}{2} K \sum_{i=1}^{18} [D - G_i(x)]^2 L_i \quad (8)$$

where nominally, $K = 0.012 \text{ N/mm/mm}$, $D = 19\text{mm}$, and L_i is the length of the weatherstrip between two points: the midpoint of points $(i-1)$ and i ; and that of points i and $(i+1)$.

Using Eq (8), the SE values for twenty seven sets of x , covering the realistic ranges of x , are generated. These are shown in Table I. Using Eq (7) and data in Table I, a regression of SE on x is then made to estimate the coefficients a_0 , b_1 , and c_{11} . The results are shown in Table II. These coefficients, together with Eq (7), completely describe the relationship between the response SE and the factors x .

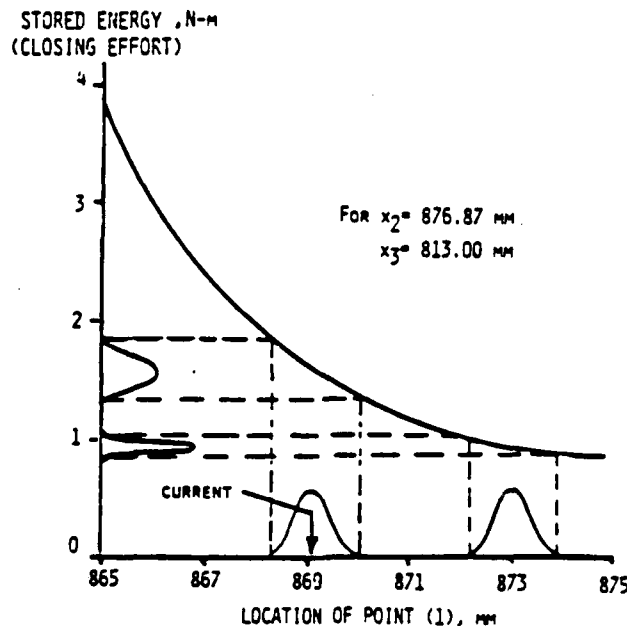


FIGURE 3. STORED ENERGY VERSUS UPPER HINGE LOCATION.

For example, Figure 3 is a plot of Eq (7) depicting how SE varies with x_1 , the position of the upper hinge. The lower hinge and the latch are fixed at their current nominal positions. The sensitivity of SE to hanging is now apparent. With x_1 at its current nominal position of 869.26 mm, a 1.0 mm deviation in x_1 produces a deviation in SE of about 0.25 N-m. By contrast, with x_1 at say 873.00 mm, the same amount of deviation in x_1 produces only a deviation in SE of about 0.1 N-m. It is this potential for desensitizing a design to sources of variation that we try to exploit: instead of trying to tighten the 1.0 mm deviation to a smaller value, we reconfigure the design, positioning x_1 at 873.00 mm, to render the design insensitive to variation.

We now exploit the full potential by considering simultaneously all three nominal positions $u = [u_1, u_2, u_3]$ of x . Avoiding drastic departure from the current positions $u^C = [869.26, 876.87, 813.00]$, we search within a 2 mm neighborhood for a position u^* at which the mean value of $SE(u^*)$ is the same as that of the current; and the variance of $SE(u^*)$ is the least. Assuming no correlations among the factors, we have:

$$\begin{aligned}\sigma_{ii} &= (0.5/3)^2, & \text{for } i = 1, 2, 3; \\ \sigma_{ij} &= 0, & \text{for } i \neq j.\end{aligned}$$

Substituting the values of σ_{ij} into Eqs (4,5,7) and simplifying, we have for a given u , the bias and variance of $SE(u)$ as follows:

$$\begin{aligned}\text{bias } [SE(u)] &= E[SE(u)] - E[SE(u^C)] \\ &= \sum_{i=1}^3 b_i (u_i - u_i^C) + \sum_{j=1}^3 \sum_{i=1}^3 c_{ij} (u_i u_j - u_i^C u_j^C); \\ \text{var } [SE(u)] &= \sum_{i=1}^3 (b_i + c_{ii} u_i + \sum_{j=1}^3 c_{ij} u_j)^2 \sigma_{ii}.\end{aligned}$$

Cast in the context of optimization, we submit the following problem to a standard optimization routine: search for u^* that

$$\begin{aligned}\text{minimizes} & \quad \sum_{i=1}^3 (b_i + c_{ii} u_i + \sum_{j=1}^3 c_{ij} u_j)^2 \sigma_{ii}; \\ \text{and satisfies} & \quad \sum_{i=1}^3 b_i (u_i - u_i^C) + \sum_{j=1}^3 \sum_{i=1}^3 c_{ij} (u_i u_j - u_i^C u_j^C) = 0; \\ & \quad u^C - 2 \text{ mm} < u^* < u^C + 2 \text{ mm}.\end{aligned}$$

The solution is $u^* = [870.27, 874.87, 811.00]$.

Figure 4 are histograms which display the results of Monte Carlo simulations of the door hanging process using the current μ^c and the optimum μ^* positions of x . For the same 0.5 mm tolerance allowed in μ , the variance of SE at optimum position is only a third of that at current position.

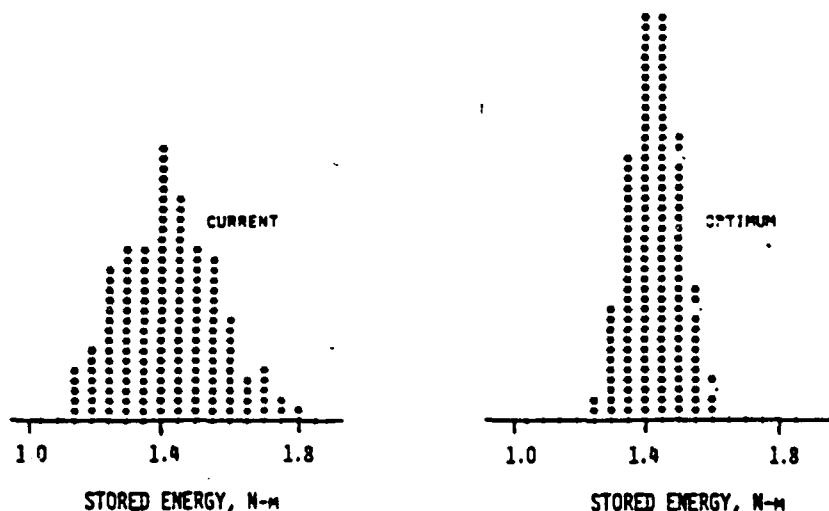


FIGURE 4. HISTOGRAMS OF STORED ENERGY (CLOSING EFFORT) FOR DOOR HANGED AT CURRENT AND AT OPTIMUM POSITION.

REFERENCES

- [1] Haftka, R.T. and Kamat, M.P., "Elements of Structural Optimization", Martinus Nijhoff Publishers, The Netherlands, 1985.
- [2] Box, G.E.P. and Fung, C.A., "Studies In Quality Improvement: Minimizing Transmitted Variation By Parameter Design", Center For Quality And Productivity Improvement Report No. 8, 1986, Madison, Wisconsin.
- [3] Hunter, J.S., "Statistical Design Applied To Product Design", Journal of Quality Technology, Vol. 17, No. 4, pp 210-221.

TABLE I - Computed Stored Energy (SE)
For Various $x = [x_1, x_2, x_3]$

x_1 (m)	x_2 (m)	x_3 (m)	<u>SE</u> (N-m)
0.86926	0.87687	0.8130	1.226
0.86926	0.87687	0.8090	2.399
0.86926	0.87687	0.8170	0.540
0.86926	0.88087	0.8130	2.745
0.86926	0.88087	0.8090	4.240
0.86926	0.88087	0.8150	2.178
0.86926	0.87587	0.8130	1.675
0.86926	0.87287	0.8090	2.024
0.86926	0.87587	0.8170	0.470
0.87226	0.87687	0.8130	0.632
0.87226	0.87687	0.8090	1.472
0.87226	0.87687	0.8140	0.498
0.87226	0.88087	0.8130	1.077
0.87226	0.88087	0.8090	2.238
0.87226	0.88087	0.8150	0.401
0.87226	0.87637	0.8130	0.679
0.87226	0.87587	0.8090	1.509
0.87126	0.87587	0.8135	0.674
0.86526	0.87687	0.8130	3.550
0.86526	0.87687	0.8090	5.168
0.86526	0.87687	0.8150	2.924
0.86526	0.87887	0.8130	4.845
0.86526	0.87887	0.8110	5.674
0.86526	0.87787	0.8150	3.485
0.86526	0.87287	0.8130	2.066
0.86526	0.87287	0.8090	3.361
0.86526	0.87287	0.8170	1.258

TABLE II - Coefficients Estimated From
The Regression of SE on x

<u>Coefficients</u>	<u>Estimate</u>
a_0 (N-m)	3821.457
b_1 (N)	-19810.583
b_2 (N)	23463.203
b_3 (N)	-13062.571
c_{11} (N/m)	42582.511
c_{22} (N/m)	40061.787
c_{33} (N/m)	7169.969
c_{12} (N/m)	-85552.756
c_{13} (N/m)	25113.789
c_{23} (N/m)	-23579.615

Session II.

Achieving Quality for
Industrial Processes

**QUALITY IMPROVEMENT: AN EXPANDING
DOMAIN FOR SCIENTIFIC METHOD**

By George Box
Center for Quality and Productivity Improvement
University of Wisconsin-Madison

QUALITY IMPROVEMENT: AN EXPANDING DOMAIN FOR SCIENTIFIC METHOD

George E. P. Box

Data Analysis and Experimental Design

Sir Ronald Fisher saw Statistics as the handmaiden of Scientific Investigation. When he first went to Rothamsted in 1919 it was to analyze data on wheat yields and rainfall that extended back over a period of some seventy years. Over the next few years Fisher developed great skill in such analysis, and his remarkable practical sense led in turn to important new theoretical results. In particular, least squares and regression analysis were placed on a much firmer theoretical basis, the importance of graphical methods was emphasized and aspects of the analysis of residuals were discussed.

But these studies presented him with a dilemma. On the one hand such "happenstance" data were affected by innumerable disturbing factors not under his control. On the other hand, if agricultural experiments were run in the laboratory, where complete control was possible, the conclusions might be valueless, since it would be impossible to know to what extent such results applied to crops grown in a farmer's field. The dilemma was resolved by Fisher's invention of statistical experimental design which did much to move science out of the laboratory and into the real world - a major step in human progress.

The theory of experimental design that he developed, employing the tools of *blocking, randomization, replication, factorial experimentation and confounding*, solved the problem of how to conduct valid experiments in a world which is naturally nonstationary

The Center for Quality and Productivity cares about your reactions to our reports. Please send comments (general or specific) to: Report Feedback, Center for Quality and Productivity Improvement, 610 Walnut Street, Madison, WI 53705. All replies will be forwarded to the authors.

and nonhomogeneous – a world, moreover, in which unseen "lurking variables" are linked in unknown ways with the variables under study, thus inviting misinformation and confusion.

Thus by the beginning of the 1930's Fisher had initiated what is now called *statistical data analysis*, had developed *statistical experimental design* and had pointed out their complementary and synergistic nature. Once the value of these ideas was demonstrated for agriculture they quickly permeated such subjects as biology, medicine, forestry and social science.

The 1930's were years of economic depression and it was not long before the potential value of statistical methods to revive and reenergize industry came to be realized. In particular at the urging of Egon Pearson and others, a new section of the Royal Statistical Society was inaugurated. During the next few years, at meetings of the Industrial and Agricultural Section of the Society, workers from academia and industry met to present and discuss applications to cotton spinning, woollen manufacture, glass making, electric light manufacture, and so forth. History has shown that these pioneers were right in their belief that statistical method provided the key to industrial progress. Unhappily their voices were not heard, a world war intervened, and it was at another time and in another country that their beliefs were proved true.

Fisher was an interested participant and frequent discussant at these industrial meetings. He wrote cordially to Shewhart, the originator at Bell Labs of quality control. He also took note of the role of sampling inspection in rejecting bad products, but he was careful to point out that the rules then in vogue for selecting inspection schemes by setting producer's and consumer's risks, could not, in his opinion, be made the basis for a theory of scientific inference (Fisher 1955). He made this point in a critical discussion of the theory of Neyman and Pearson whose "errors of the first and second kind" closely paralleled producer's and consumer's risks. He could not have foreseen that fifty years later the world of *quality control* would, in the hands of the Japanese, have become the world of

quality improvement in which his ideas for scientific advance using statistical techniques of design as well as analysis were employed by industry on the widest possible basis.

The revolutionary shift from quality control to quality improvement which had been initiated in Japan by his long-time friend Dr. W. Edwards Deming was accompanied by two important concomitant changes; involvement of the whole workforce in quality improvement and recognition that quality improvement must be a continuous and never-ending occupation (see Deming (1986)).

Aspects of Scientific Method

To understand better the logic of these changes it is necessary to consider certain aspects of scientific method. Among living things mankind has the almost unique ability of making discoveries and putting them to use. But until comparatively recently such technical advance was slow – the ships of the thirteenth century were perhaps somewhat better designed than those of the twelfth century but the differences were not very dramatic. And then three or four hundred years ago a process of quickened technical change began which has ever since been accelerating. This acceleration is attributed to an improved process for finding things out which we call *scientific method*.

We can, I think, explain at least some aspects of this scientific revolution by considering a particular instance of discovery. We are told that in the late seventeenth century it was a monk from the Abbey of Hautvillers who first observed that a second fermentation in wine could be induced which produced a new and different sparkling liquid, delightful to the taste, which we now call champagne. Now the culture of wine itself is known from the earliest records of man and the conditions necessary to induce the production of champagne must have occurred countless times throughout antiquity. However, it was not until this comparatively recent date that the actual discovery was made. This is less surprising if we consider that to induce an advance of this kind two

circumstances must coincide. First an *informative event* must occur and second a *perceptive observer* must be present to see it and learn from it.

Now most events that occur in our daily routine correspond more or less with what we expect. Only occasionally does something occur which is potentially informative. Also many observers, whether through lack of essential background knowledge or from lack of curiosity or motivation, do not fill the role of a perceptive observer.

Thus the slowness in antiquity of the process of discovery can be explained by the extreme rarity of the chance *coincidence* of two circumstances *each* of which is itself rare. It is then easily seen that discovery may be accelerated by two processes which I will call informed observation and directed experimentation.

By a process of *informed observation* we arrange things so that, when a rare potentially informative event *does* occur people with necessary technical background and motivation are there to observe it. Thus, when last year an explosion of a supernova occurred, the scientific organization of this planet was such that astronomers observed it and learned from it. A quality control chart fills a similar role. When such a chart is properly maintained and displayed it ensures that any abnormality in the routine operation of a process is likely to be observed and associated with what Shewhart called an *assignable cause* – so leading to the gradual elimination of disturbing factors.

A second way in which the rate of acquisition of knowledge may be increased is by what I will call *directed experimentation*. This is an attempt to artificially induce the occurrence of an informative event. Thus, Benjamin Franklin's plan to determine the possible connection of lightning and electricity by flying a kite in a thunder cloud and testing the emanations flowing down the string, was an invitation for such an informative event to occur.

Recognition of the enormous power of their methods of scientific advance is now commonplace. The challenge of the modern movement of quality improvement is nothing less than to use them, to further in the widest possible manner the effectiveness of human

activity. By this I mean not only the process of industrial manufacture but the running, for example, of hospitals, airlines, universities, bus services and supermarkets.

The enormous potential of such an approach had long been foreseen by systems engineers (see for example, Jenkins & Youle (1971)) but until the current demonstration by Japan of its practicability, it had been largely ignored.

Organization of Quality Improvement with Simple Tools

The less sophisticated problems in quality improvement can often be solved by informed observation using some very simple tools that are easily taught to the workforce.

While on the one hand, Murphy's law implacably ensures that anything that can go wrong with a process will eventually go wrong, this same law also ensures that every process produces data which can be used for its own improvement. In this sentence the word process could mean an industrial manufacturing process, or a process for ordering supplies or for paying bills. It could also mean the process of admission to a hospital or of registering at a hotel or of booking an airline flight.

One major difficulty in past methods of system design was the lack of involvement of the people closest to it. For instance, a friend of mine recently told me of the following three incidents that happened on one particular day. In the morning he saw his doctor at the hospital to discuss the results of some tests that had been made two weeks before. The results of the tests should have been entered in his records but, as frequently happened at this particular hospital, they were not. The doctor smiled and said rather triumphantly, "Don't worry, I thought they wouldn't be in there. I keep a duplicate record myself although I'm not supposed to. So I can tell you what the results of your tests are." Later that day my friend flew from Chicago to New York and as the plane was taxiing prior to takeoff there was a loud scraping noise at the rear of the plane. Some passengers looked concerned but said nothing. My friend pressed the call button and asked the stewardess about it. She said, "This plane always makes that noise but obviously I can't do anything

about it." Finally on reaching his hotel in the evening he found that his room, the reservation for which had been guaranteed, had been given to someone else. He was told that he would be driven to another hotel and that because of the inconvenience his room-rate would be reduced. In answer to his protest the reservation clerk said, "I'm very sorry -- it's the system. It's nothing to do with me."

In each of these examples the system was itself providing data which could have been used to improve it. But in every case improvement was frustrated, because the doctor, the stewardess and the hotel clerk each believed that there was nothing they could do to alter a process that was clearly faulty. Yet each of the people involved was much closer to the system than those who had designed it and who were insulated from receiving data on how it could be improved.

Improvement could have resulted if, in each case, a routine had been in place whereby data coming from the system were automatically used to correct it.

To achieve this it would first have been necessary

- (a) to instill the idea that quality improvement was each individual persons responsibility,
- (b) to move responsibility for the improvement of the system to a quality improvement team which included the persons actually involved,
- (c) to organize collection of appropriate data (not as a means of apportioning blame, but to provide material for team problem-solving meetings).

The quality team of the hospital records system might include the doctor, the nurse, someone from the hospital laboratory and someone from the records office. For the airplane problem, the team might include the stewardess, the captain, and the person responsible for the mechanical maintenance of the plane. For the hotel problem, the team might include the hotel clerk, the reservations clerk, and someone responsible for computer systems. It is the responsibility of such teams to conduct a relentless and never-ending war against Murphy's regime. Because their studies often reveal the need for

additional expertise, and because it will not always be within the power of the team to institute appropriate corrective action, it is necessary that adequate channels for communication exist from the bottom to the top of the organization, as well as from the top to the bottom.

Three potent weapons in the campaign for quality improvement are *Corrective Feedback*, *Preemptive Feedforward* and *Simplification*. The first two are self explanatory. The importance of simplification has been emphasized and well illustrated by Tim Fuller (1986). In the past systems have tended to become steadily more complicated without necessarily adding any corresponding increase in effectiveness. This occurs when

- (a) the system develops by reaction to occasional disasters,
- (b) action, supposedly corrective, is instituted by persons remote from the system,
- (c) no check is made on whether corrective action is effective or not.

Thus the institution by a department store of complicated safeguards in its system for customer return of unsatisfactory goods might be counter-productive; while not providing sufficient deterrence to a mendacious few, it could cause frustration and rejection by a large number of honest customers. By contrast the success of a company such as Marks & Spencer who believe instead in simplification and, in particular, adopt a very enlightened policy toward returned goods, speaks for itself.

Because complication provides work and power for bureaucrats, simplification must be in the hands of people who can benefit from it. The time and money saved from quality improvement programs of this sort far more than compensates for that spent in putting them into effect. No less important is the great boost to the morale of the workforce that comes from their knowing that they can use their creativity to improve efficiency and reduce frustration.

Essential to the institution of quality improvement is the redefinition of the role of the manager. He should not be an officer who conceives, gives and enforces orders but rather a coach who encourages and facilitates the work of his quality teams.

Ishikawa's Seven Tools

At a slightly more sophisticated level the process of informed observation may be facilitated by a suitable set of statistical aids typified, for example, by Ishikawa's seven tools. They are described in an invaluable book available in English (Ishikawa 1976) and written for foremen and workers to study together. The tools are *check sheets*, *Pareto charts*, *cause-effect diagrams*, *histograms*, *graphs*, *stratification* and *scatter plots*. They can be used for the study of service systems as well as manufacturing systems but I will use an example of the latter kind (see, for example, Box and Bisgaard, (1987)).

Suppose a manufacturer of springs finds that at the end of a week that 75 springs have been rejected as defective. These rejects should not be thrown away but studied by a quality team of the people who make them. As necessary this team would be augmented from time to time with appropriate specialists. A tally on a check sheet could categorize the rejected springs by the nature of the defect. Display of these results on a Pareto chart might then reveal the primary defect to be, say, cracks. To facilitate discussion of what might cause the cracks the members of the quality team would gather around a blackboard and clarify their ideas using a cause-effect diagram. A histogram categorizing cracks by size would provide a clear picture of the magnitude of the cracks of how much they varied. This histogram might then be stratified, for example, by spring type. A distributional difference would raise the question as to why the cracking process affected the two kinds of springs differently and might supply important clues as to the cause. A scatter plot could expose a possible correlation of crack size with holding temperature and so forth. With simple tools of this kind the team can work as *quality detectives* gradually "finding and fixing" things that are wrong.

It is sometimes asked if such methods work outside Japan. One of many instances showing that it can, is supplied by a well-known Japanese manufacturer making television sets just outside Chicago in the United States. The plant was originally operated by an American company using traditional methods of manufacture. When the Japanese

company first took over the reject rate was 146%. This meant that most of the television sets had to be taken off the line once to be individually repaired and some had to be taken off twice. By using simple "find and fix" tools like those above the reject rate over a period of 4 - 5 years was reduced from 146% to 2%. Although this was a Japanese company only Americans were employed, and a visitor could readily ascertain that they greatly preferred the new system.

Evolutionary Operation

Evolutionary Operation (EVOP) is an example of how elementary ideas of *experimental design* can be used by the whole workforce. The central theme (Box 1957, Box and Draper 1969) is that an operating system can be organized which mimics that by which biological systems evolve to optimal forms.

For manufacturing, let us say, a chemical intermediate, the standard procedure is to continually run the process at fixed levels of the process conditions—temperature, flow rate, pressure, and agitation speed and so forth. Such a procedure may be called *static operation*. However, experience shows that the best conditions for the full scale process are almost always somewhat different from those developed from smaller scale experimentation and furthermore, that some factors important on the full scale cannot always be adequately simulated in smaller scale production. The philosophy of Evolutionary Operation is that the full scale process may be run to produce not only product, but also information on how to improve the process and the product. Suppose that temperature and flow rate are the factors chosen for initial study. In the *evolutionary operation* mode small deliberate changes could be made in these two factors in a pattern (an experimental design) about the current best-known conditions. By continuous averaging and comparison of results at the slightly different conditions as they come in, information gradually accumulates which can point to a direction of improvement where for example higher conversion or less impurity can be obtained.

Important aspects of Evolutionary Operation are

- (a) it is an alternative method of continuous process operation. It may therefore be run indefinitely as new ideas evolve and the importance of new factors are realized.
- (b) it is run by plant operators as a standard routine with the guidance of the process superintendent and occasional advice from an evolutionary operation committee. It is thus very sparing in the use of technical manpower.
- (c) it was designed for improving yields and reducing costs in the chemical and process industries. In the parts industries, where the problem is often that of reducing variation by studying variances instead of means at the various process conditions the process can be made to evolve to one where variation is minimized.

Design of Experiments for Engineers

The methods described so far are ways of doing the best we can with what we have, assuming that the basic design of the product we produce and the process that produces it are essentially immutable. Obviously a product or process, which suffers from major deficiencies of design, cannot be improved beyond a certain point by these methods.

However by artful design of a new product and of the process that makes it, it may be possible to arrive both at a highly efficient process of manufacture and a product that behaves well and almost never goes wrong. The design of new products and processes is a fertile field for the employment of statistical experimental design by engineers.

Which? How? Why?

Suppose y is some quality characteristic whose probability distribution depends on the level of a number of factors x . Experimental design may be used to reveal certain aspects of this dependence; in particular how the mean $E(y) = f(x)$, and the variance

$\sigma^2(y) = F(x)$, depend on x . Both choice of design and method of analysis are greatly affected by what we know or what we think we know about the input variables x and the functions $f(x)$ and $F(x)$ (see for example Box, Hunter and Hunter (1978)).

Which: In the early stages of investigation the task may be to determine *which* subset of variables x_k chosen from the larger set (x) are of importance in affecting y .

In this connection a Pareto hypothesis (a hypothesis of "effect sparsity") becomes appropriate and the projective properties into lower dimensions in the factor space of highly fractionated designs may be used to find an active subset of k or fewer active factors. Analyses based on normal plots and/or Bayesian methods (Box and Meyer 1986a)) are efficient and geometrically appealing.

How: When we know or think we know which are the important variables x_k we may need to determine more precisely how changes in their levels affect y . Often the nature of the functions $f(x)$ and $F(x)$ will be unknown. However over some limited region of interest a local Taylor's series approximation of first or second order in x_k may provide an adequate approximation, particularly if y and x_k may be re-expressed when necessary in appropriate alternative metrics. Fractional factorials and other response surface designs of first and second order are appropriate here. Maxima may be found and exploited using steepest ascent methods followed by canonical analysis of a fitted second degree equation in appropriately transformed metrics. The possibilities for exploiting multidimensional ridges and hence alternative optimal process become particularly important at this stage (see for example Box and Draper 1986).

Why: Instances occur when a mechanistic model can be postulated. This might take the form of a set of differential equations believed to describe the underlying physics.

Various kinds of problems arise. Among these are:

How should the parameters (often corresponding to unknown physical constants) be estimated from the data?

How should candidate models be tested?

How should we select a model from competing candidates?

What kinds of experimental designs are appropriate?

Workers in quality improvement have so far been chiefly occupied with problems of the "which" and occasionally of the "how" kind and have consequently made most use of fractional factorial designs, other orthogonal arrays, and response surface designs.

Studying Location, Dispersion and Robustness

In the past experimental design had been used most often as a means of discovering how x_k affected the mean value $E(y)$ how, for example, the process could be improved by increasing the mean of some quality characteristics. Modern quality improvement also stresses the use of experimental design in reducing dispersion as measured, for example, by the variance.

Using experimental designs to minimize variation: High quality particularly in the parts industries (e.g. automobiles, electronics) is frequently associated with minimizing dispersion. In particular the simultaneous study of the effect of the variables x on the mean and on the variance is important in the problem of bringing a process on target with smallest possible dispersion (Phadke 1982). Bartlett and Kendall (1946) pointed to the advantages of analysis in terms of $\log s_y^2$ to produce constant variance and increased additivity in the dispersion measure. It is also very important in such studies to remove transformable dependence between the mean and standard deviation. Taguchi (1986,1987) attempts to do this by the use of a signal to noise ratios. However, it may be shown that it is much less restrictive, simpler and more statistically efficient to proceed by direct data transformation obtained, for example, by a "lambda plot" (Box 1988).

A practical difficulty may be the very large number of experimental runs which may be needed in such studies if complicated designs are employed. It is recently shown how using what Fisher called hidden replication, unreplicated fractions may sometimes be

employed to identify sparse dispersion effects in the presence of sparse location effects (Box and Meyer 1986b)).

Experimental Design and Robustness to the Environment: A well designed car will start over a wide range of conditions of ambient temperature and humidity. The design of the starting mechanism may be said to be "robust" to changes in these environmental variables. Suppose $E(y)$ and possibly also $\sigma^2(y)$ are functions of certain design variables x_d which determine the design of the system and also of some external environmental variables x_v which, except in the experimental environment, are not under our control. The problem of robust design is to choose a desirable combination of design variables x_{do} at which good performance is experienced over a wide range of environmental conditions.

Related problems were earlier considered by Youden (1961a,b) and Wernimont (1975) but recently their great importance in quality improvement has been pointed out by Taguchi. His solution employs an experimental design which combines multiplicatively an "inner" design array and an "outer" environmental array. Each piece of this combination is usually a fractional factorial design or some other orthogonal array. Recent research has concentrated on various means for reducing the burdensome experimental effort which presently may be needed for studies of this kind.

Robustness of an assembly to variation in its components: In the design of an assembly, such as an electrical circuit, the exact mathematical relation $y = f(x)$ between the quality characteristic of the assembly, such as the output voltage y of the circuit, and the characteristics x of its components (resistors, capacitors, etc.) may be known from physics. However there may be an infinite variety of configurations of x that can give the same desired mean level $E(y) = \eta$, say. Thus an opportunity exists for optimal design by choosing a "best" configuration.

Suppose the characteristics x of the components vary about "nominal values" ξ with known covariance matrix V . Thus for example a particular resistance x_i might vary

about its nominal value ξ_i with known variance σ_i^2 . (Also variation in one component would usually be independent of that of another so that V would usually be diagonal.)

Now variation in the input characteristics x will transmit variation to the quality characteristic y so that for each choice of component nominal values ξ which yield the desired output $y = \eta$ there will be an associated mean square error $E(y - \eta)^2 = M(\eta) = F(\xi)$.

Using a Wheatstone Bridge circuit for illustration, Taguchi and Wu (1985) pose the problem of choosing ξ so that $M(\eta)$ is minimized. To solve it they again employ an experimental strategy using inner and outer arrays. Box and Fung (1986) have pointed out, however, that their procedure does not in general lead to an optimal solution and that it is better to use a simpler and more general method employing a standard numerical nonlinear optimization routine. The latter authors also make the following further points.

- (a) For an electrical circuit it is reasonable to assume that the relation $y = f(x)$ is known, but when, as is usually the case, $y = f(x)$ must be estimated experimentally, the problems are much more complicated and require further study.
- (b) It is also supposed that each of the σ_i^2 are known and furthermore that they *remain fixed or change in a known way* (for example proportionally) when ξ_i changes. The nature of the optimal solution can be vastly different depending on the validity of such assumptions.

Taguchi's quality engineering ideas are clearly important and present a great opportunity for development. It appears however (see, for example, Box, Bisgaard and Fung (1988)) the accompanying statistical methods that Taguchi recommends employing "accumulation analysis," "signal to noise ratios" and "minute analysis" are often defective, inefficient and unnecessarily complicated. Furthermore, Taguchi's philosophy seems at times to imply a *substitution* of statistics for engineering rather than the use of statistics as

a *catalyst* to engineering (Box 1988). Because such deficiencies can be easily corrected it is particularly unfortunate that, in the United States at least, engineers are often taught these ideas by instructors who stress that no deviation from Taguchi's exact recipe is permissible.

A Wider Domain for Scientific Method

Quality improvement is about finding out how to do things better. The efficient way to do this is by using scientific method—a very powerful tool, employed in the past by only a small elite of trained scientists. Modern quality improvement extends the domain of scientific method over a number of dimensions:

over *users* (e.g. from the chief executive officer to the janitor)

over *areas* of human endeavor (e.g. factories, hospitals, airlines, department stores)

over *time* (never-ending quality improvement)

over *causative factors* (an evolving panorama of factors that effect the operation of a system).

Users: Although it is not possible to be numerically precise I find a rough graphical picture helpful to understanding. The distribution of technological skill in the workforce might look something like Figure 1(a). The distribution of technological skill required to solve the problems, that routinely reduce the efficiency of factories, hospitals, bus companies and so forth, might look something like Figure 1(b).

In the past only those possessing highly trained scientific or managerial talent, would have been regarded as problem solvers. Inevitably this small group could only tackle a small proportion of the problems that beset the organization. One aspect of the new approach is that many problems can be solved by suitably trained persons of lesser technical skills. An organization that does not use this talent throws away a large proportion of its creative potential. A second aspect is that engineers and technologists must be taught how to experiment simultaneously with many variables in the presence of noise. Without knowledge of statistical experimental design they are not equipped to do

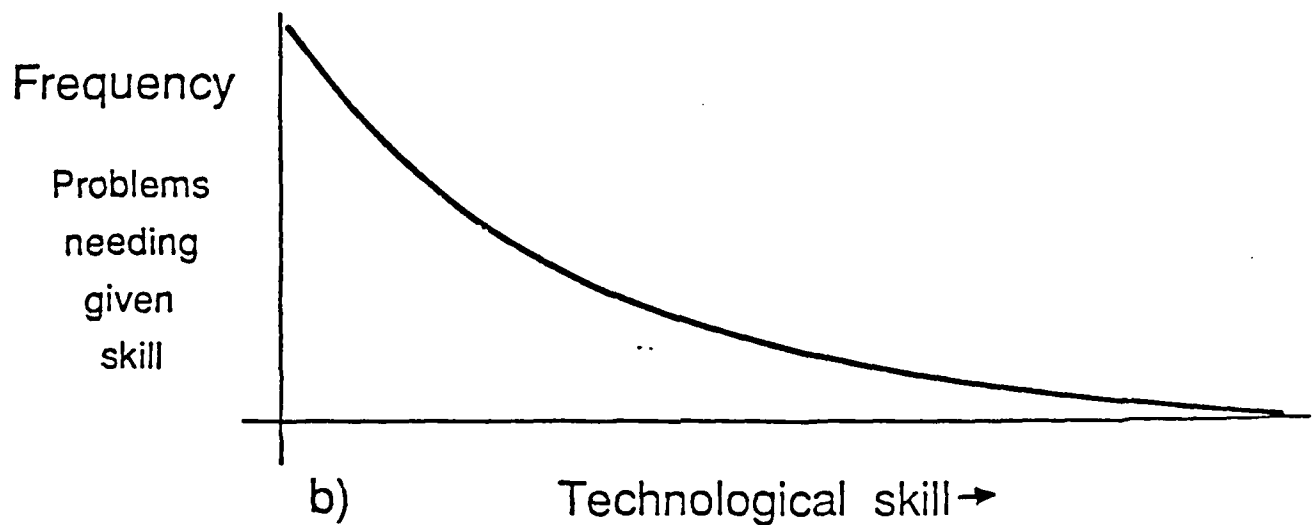
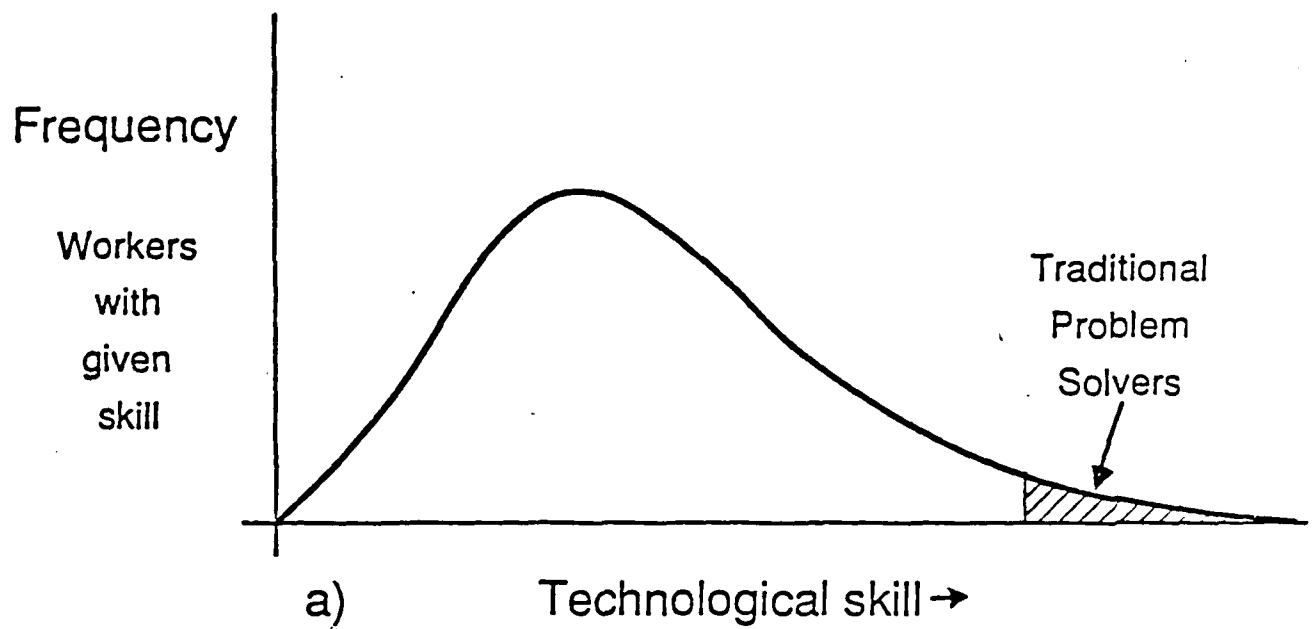


Figure 1. 1a) Hypothetical frequency distribution of workers possessing given technical skill
 1b) Hypothetical frequency distribution of problems requiring various design of technological skill for their solution

this. A group visiting Japanese industry was recently told "an engineer who does not know statistical experimental design is not an engineer." (Box et al 1988.)

Areas of Endeavor: At first sight we tend to think of quality improvement as applying only to operations on the factory floor. But even in manufacturing organizations a high proportion of the workforce are otherwise engaged – in billing, invoicing, planning, scheduling and so forth – all of which should be made the subject of study.

But the individual citizen in every day of his life must deal with an unnecessarily complex world, involving hospitals, government departments, universities, airlines and so forth. Lack of quality in these organizations results in needless expense, wasted time and unnecessary frustration. Quality improvement applied to these activities could free us all for more productive and pleasurable pursuits.

Time: For never ending improvement there must be a long-term commitment to renewal. A commonly used statistical model links a set of variables x_k with a response y by an equation $y = f(x_k) + e$ where e is an error term, often imbued by statisticians with properties of randomness, independence and normality. A more realistic version of this model is

$$y = f(x_k) + e(x_u)$$

where x_u is a set of variables whose nature and behavior is unknown. As time elapses, by skillful use of the techniques of informed observation and experimental design, as time elapses, elements of x_u are transferred into x_k — from the unknown into the known.

This transference is the essence of modern quality improvement and has two consequences:

- (a) once a previously unknown variable has been identified it can be fixed at a level that produces best results.
- (b) by fixing it we remove an element previously contributing to variation.

This transfer can be never-ending process whereby knowledge increases and variation is reduced.

The structure of the process of investigation so far as it involves statistics has not always been understood. It has sometimes been supposed that it consists of testing a null hypothesis selected in advance against alternatives using a single set of data. In fact most investigations proceed in an iterative fashion in which *deduction* and *induction* proceed in alternation (see, for example, Box (1980)). Clearly optimization of individual steps in such a process can lead to sub-optimization for the investigational process itself. The inferential process of *estimation* whereby a postulated model and supposedly relevant data are *combined* is purely *deductive* and resembles a process of "summation" of model and data. It is conditional on the assumption that the model of the data generating process and the actual process itself are consonant. No warning that they are not consonant is provided by estimation. However a comparison of appropriate qualities derived from the data with a sampling reference distribution generated by the model provides a process of criticism that can not only discredit the model but suggest appropriate direction for model modification. An elementary example of this is Shewhart's "assignable cause" deduced from *data* outside control lines which are calculated from a *model* of the data-generating process *in a state of control*. Such a process of *criticism contrasts* features of the model and the data and thus resembles a process of "differencing" of model and data. It can lead the engineer, scientist or technologist by a process of *induction* to postulate a modified, or a totally different model, so recharting the course for further exploration. This process is subjective and artistic. It is the only step that can introduce new ideas and hence must be encouraged above all else. It is best encouraged, I believe, by interactive graphical analysis. This is readily provided these days by computers which also make it possible to use sophisticated statistical ideas that are calculation-intensive and yet produce simply understood graphical output. It is by following such a deductive-inductive iteration that the quality investigator can be led to a solution of a problem just as a good detective can solve a murder mystery.

Factors and Assignable Causes: The field of factors potentially important to quality improvement also can undergo seemingly endless expansion. Problems of optimization are frequently posed as that of maximizing some response y over a k -dimensional space of known factors x_k , but in quality improvement the factor space is never totally known and is continually developing.

Consider a possible scenario for a problem which begins as that of choosing the reaction time x_1 and reaction temperature x_2 to give maximum conversion y of raw materials to the desired product. Suppose experimentation with these two factors leads to the (conditional) optimal choice of coordinates in Figure 2(a). Since conversion is only 50% we see that the best is not very good if we restrict the system in this way. After some deliberation it is now suggested that higher conversion might be obtained if biproducts which may be being formed at the beginning of the reaction were suppressed by employing a lower temperature in the early stages. This idea produces two new variables, the initial temperature x_3 and the time x_4 taken to reach the final temperature. Their best values, and the appropriately changed levels of x_1 and x_2 , might be those shown in Figure 2(b). This new (conditional) optimal profile again results in only partial success ($y = 68\%$) leading to the suggestion that the (newly increased) final temperature may result in other biproducts being formed at the *later* stages of reaction. This suggests experimentation with variables x_5 and x_6 allowing for a fall off in temperature towards the end of the reaction. The new (conditional) optimal profile at this stage might then be as in Figure 2(c).

These results might now be seen by a physical chemist leading him to suggest a mechanistic theory yielding a series of curved profiles which depended on *only two* theoretical constants X_1 and X_2 . If this idea was successful the introduction of two new experimental factors would have eliminated the need for the other six. The new mechanistic theory could in turn suggest new factors, not previously thought of, which might produce even greater conversion and so on. Thus the factor space must realistically

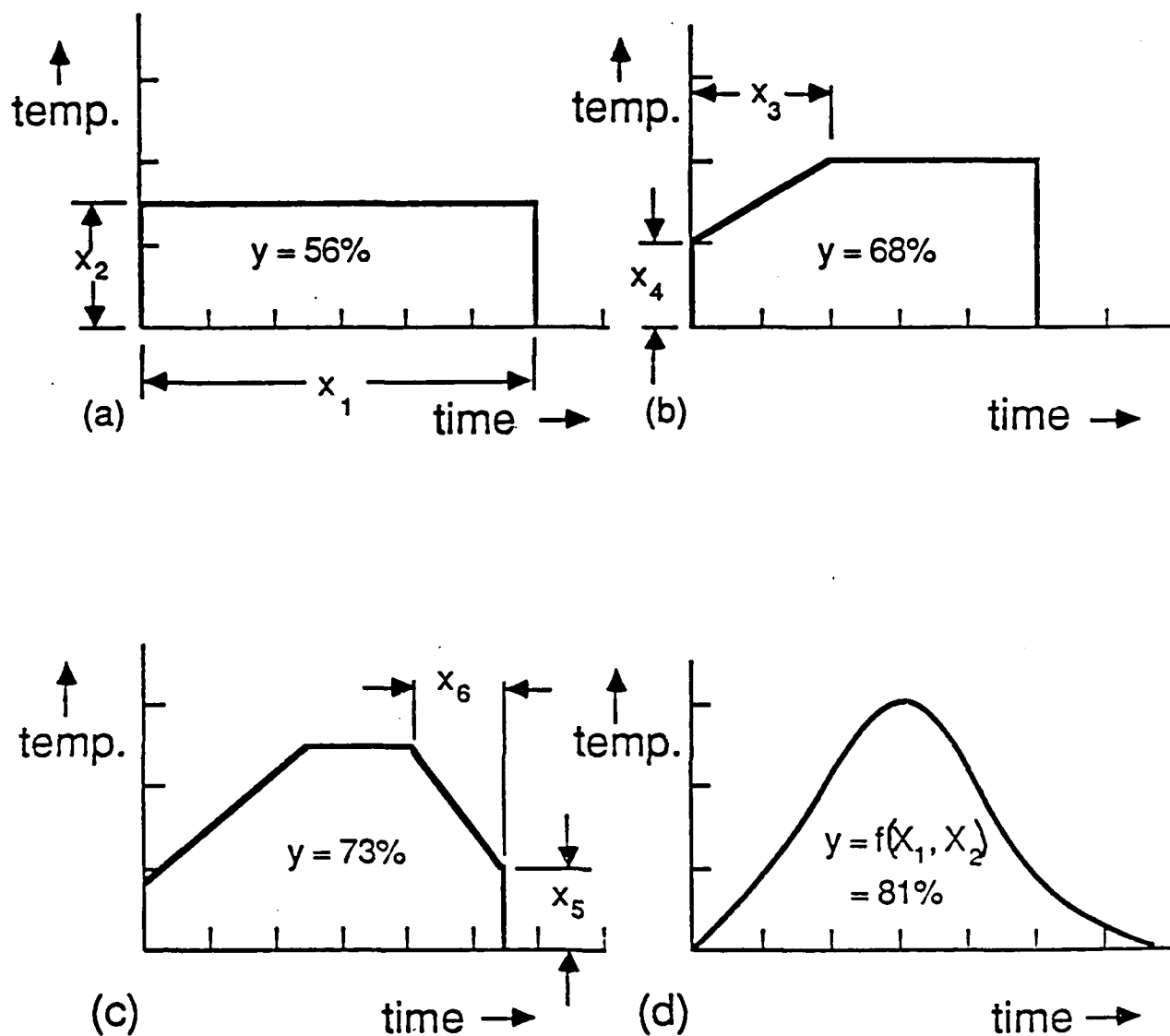


Figure 2. Development of the factor space in a problem of maximizing experimentally the chemical conversion y as a function of time and temperature

be regarded as one which is continually changing and developing as part of the evolution which occurs within the scientific process in a manner discussed earlier by Kuhn (1962).

Training

Instituting the necessary training for quality is a huge and complex task. Some assessment must be made of the training needs for the workforce, for engineers, technologists and scientists, and for managers at various levels, and we must consider how such training programs can be organized using the structure that we have within industry, service organizations, technical colleges and universities.

A maximum multiplication effect will be achieved by a scheme in which the scarce talent that is available is employed to teach the teachers within industry and elsewhere.

It is, I believe, unfortunately true that the number of graduates who are interested in industrial statistics in Great Britain has been steadily decreasing. The reasons for this are complex and careful analysis and discussion between industry, the government, and the statistical fraternity is necessary to discover what might be done to rectify the situation.

Management

The case for the extension of scientific method to human activities is so strong and so potentially beneficial that one may wonder why this revolution has not already come about. A major difficulty is to persuade the managers.

In the United States it is frequently true that both higher management and the workforce are in favor of these ideas. Chief executive officers whose companies are threatened with extinction by foreign competition are readily convinced and are prepared to exhort their employees to engage in quality improvement. This is certainly a step in the right direction but it is not enough. I recently saw a poster issued jointly by the Union and the Management of a large automobile company in the United States which reiterated a Chinese proverb "Tell me — I'll forget; Show me — I may remember; Involve me and

I'll understand". I'm not sure that top management always realize that their involvement (not just their encouragement) is essential. The workforce enjoy taking part in quality improvement and will cooperate provided they believe that the improvements they bring about will not be used against them.

Some members of the middle levels of management and of the bureaucracy pose a more serious problem because they see loss of power in sharing the organization of quality improvement with others. Clearly the problems of which Mr. Gorbachev complains in applying his principles of perestroika are not confined to the Soviet Union.

Thus the most important questions are really not concerned so much with details of the techniques but with whether a process of change in management can be brought about so that they can be used at all. It is for this reason that Dr. Deming and his followers have struggled so hard with this most difficult problem of all — that of inducing the changes in management and instituting the necessary training that can allow the powerful tools of scientific method to be used for quality improvement. It is on the outcome of this struggle that our economic future ultimately depends.

Fisher's Legacy

At the beginning of this lecture I portrayed Fisher as a man who developed statistical methods of design and analysis that extended the domain of science from the laboratory to the whole world of human endeavor. We have the opportunity now to bring that process to full fruition. If we can do this we can not only look forward to a rosier economic future, but by making our institutions easier to deal with, we can improve the quality of our everyday lives, and most important of all we can joyfully experience the creativity which is a part of every one of us.

Acknowledgement

This research was sponsored by the National Science Foundation Grant No. DMS-8420968, the United States Army under Contract NO. DAAL03-87-K-0050, and by the Vilas Trust of the University of Wisconsin-Madison.

References

- Bartlett, M.S. and D.G. Kendall (1946) The statistical analysis of variance - heterogeneity and the logarithmic transformation. *J. Roy. Statist. Soc.*, Ser B 8, 128-150.
- Box, G.E.P. (1957) Evolutionary Operation: A method for increasing industrial productivity. *Applied Statistics* 6, 3-23.
- Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist.Soc.*, Ser A, 143, 383-404.
- Box, G.E.P. (1988) Signal-to-noise ratios, performance criteria and transformations. *Technometrics* 30, 1-17.
- Box, G.E.P. and S. Bisgaard (1987) The scientific context of quality improvement. *Quality Progress*, 20, No. 6, 54-61.
- Box, G.E.P., Bisgaard S., and C.A. Fung (1988) An explanation and critique of Taguchi's contributions to quality engineering. Report No. 28, Center for Quality and Productivity Improvement, University of Wisconsin-Madison. To appear: *Quality and Reliability Engineering International*, May 1988.
- Box, G.E.P. and N.R. Draper (1969) Evolutionary Operation—A statistical method for process improvement. New York: Wiley.
- Box, G.E.P. and N.R. Draper (1986) Empirical model-building and response surfaces. New York: Wiley.

- Box, G.E.P. and C. A. Fung (1986) Studies in quality improvement: Minimizing transmitted variation by parameter design. Report No. 8, Center for Quality and Productivity Improvement, University of Wisconsin-Madison.
- Box, G.E.P., Hunter, W.G., and J.S. Hunter (1978) Statistics for experimenters: An introduction to design, data analysis, and model building. New York: Wiley.
- Box, G.E.P. and R.D. Meyer (1986a) An analysis for unreplicated fractional factorials. *Technometrics* 28, 11-18.
- Box, G.E.P. and R.D. Meyer (1986b) Dispersion effects from fractional designs. *Technometrics* 28, 19-27.
- Box, G.E.P., Kacker, R.N., Nair, V.N., Phadke, M., Shoemaker, A.C., and C.F. J. Wu (1988) Quality Practices in Japan. *Quality Progress*, XX, No. 3, pp 37-41.
- Deming, W.E. (1986) Out of the crisis, Cambridge MA, MIT Press.
- Fisher, R.A. (1955) Statistical methods and scientific induction. *J. Roy. Statist. Soc.*, Ser. B, 17, 69-78
- Fuller, F.T. (1986) Eliminating complexity from work: improving productivity by enhancing quality. Report No. 17, Center for Quality and Productivity Improvement, University of Wisconsin-Madison.
- Ishikawa, K. (1976) Guide to quality control. Asian Productivity Organization.
- Jenkins, G.M. and P.V. Youle (1971) Systems engineering. London; G.A. Watts.
- Kuhn, T. (1962) The Structure of Scientific Revolutions. *University of Chicago Press*, Chicago.
- Phadke, M.S. (1982) Quality engineering using design of experiments. Proceedings of the section on Statistical Education, American Statistical Association, 11-20.
- Taguchi, G. (1986) Introduction to quality engineering: Designing quality into products and processes. White Plains, N.Y.: *Kraus International Publications*.
- Taguchi, G. (1987) System of experimental design . Vol 1 and 2, White Plains, NY: *Kraus International Publications*.

- Taguchi, G. and Y. Wu (1985) Introduction to off-line quality control. Nagaya Japan: Central Japan Quality Control Association.
- Wernimont, G. (1977) Ruggedness evaluation of test procedures. *Standardization News*, 5, 13-16.
- Youden, W.J. (1961a) Experimental design and ASTM committees. *Materials research and standards*, 1, 862-867, Reprinted in *Precision Measurement and Calibration*, Special Publication 300, National Bureau of Standards, Vol. 1, 1969, Editor H. H. Ku.
- Youden, W.J. (1961b) Physical measurement and experimental design. *Colloques Internationaux du Centre National de la Recherche Scientifique No 110, le Plan d'Experiences*, 1961, 115-128. Reprinted in *Precision Measurement and Calibration*, Special Publication 300, National Bureau of Standards, Vol. 1, 1969, Editor H. H. Ku.

QUALITY IMPROVEMENT APPROACHES FOR
CHEMICAL PROCESSES

William J. Hill
Allied-Signal Inc.

Presentation at NBS Conference on
"Uncertainty in Engineering Design",
Gaithersburg, MD May 10-11, 1988

6/4/82

"SOMETIMES A PARTICULAR ISSUE IS STUDIED BY A NUMBER OF DIFFERENT INVESTIGATORS WITH MIXED RESULTS SOME OF WHICH MAY EVEN SEEM TO BE CONTRADICTIONARY. BUT IT IS IMPOSSIBLE THAT RESULTS BE CONTRADICTIONARY IN ANY ABSOLUTE SENSE BECAUSE NATURE MARCHES TO BUT A SINGLE DRUMMER. TRUE, SOME OF THE DRUM ROLLS MAY BE INTRICATE AND, BECAUSE OF DEFECTS IN OUR HEARING, WE MAY NOT DISCERN ALL THE NOTES. NATURE, HOWEVER, NEVER MISSES A BEAT. SHE IS A PERFECT MARCHER. THE NOTES WE MISS ARE THE LURKING VARIABLES. IN SORTING OUT QUESTIONS OF CAUSE AND EFFECT, THE JOB OF SCIENCE IS TO FIGURE OUT HOW THE DRUMMER CALLS THE TUNE.

IF WE SEE SOME SEEMINGLY ERRATIC MARCHING, THE FAULT LIES NOT IN THE MARCHER BUT IN OUR INABILITY TO INFER THE MATCH OF MUSIC TO MARCH, MOST OFTEN BECAUSE WE DO NOT HEAR ALL THE NOTES."

BILL HUNTER

In Chemical Process Industry

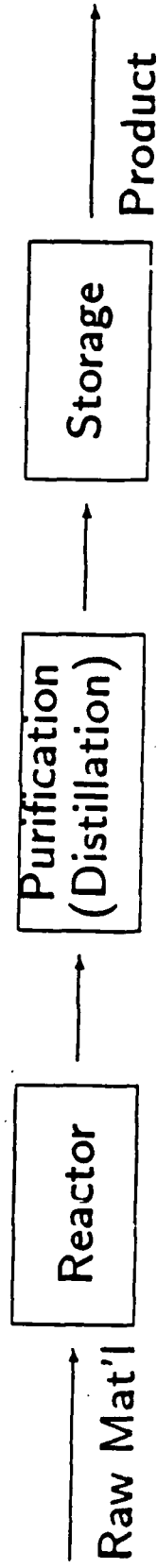
We deal in:

1) Liquids

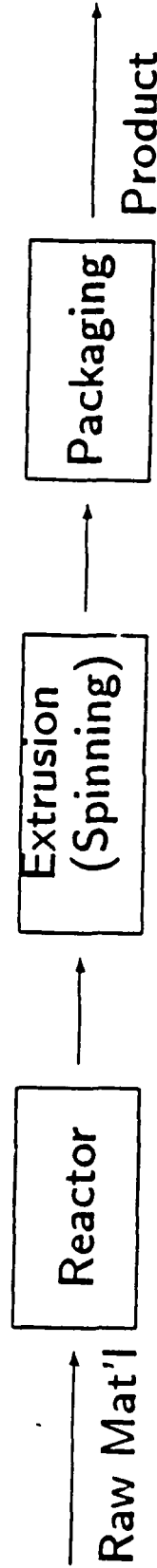
2) Gases

3) Extruded Solids
e.g., Nylon
Plastic

Liquids/Gases



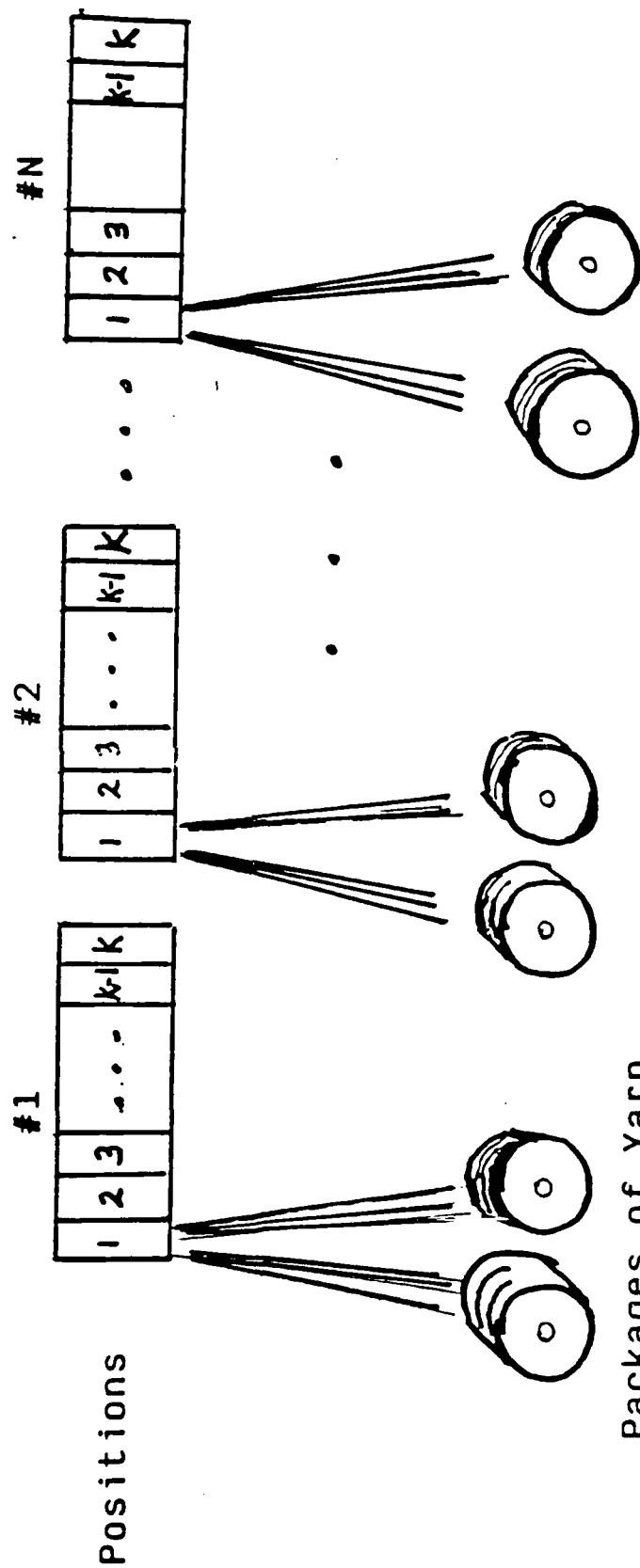
Nylon/Plastic



Further Complications

Expansion: Extrusion/Spinning

Machines

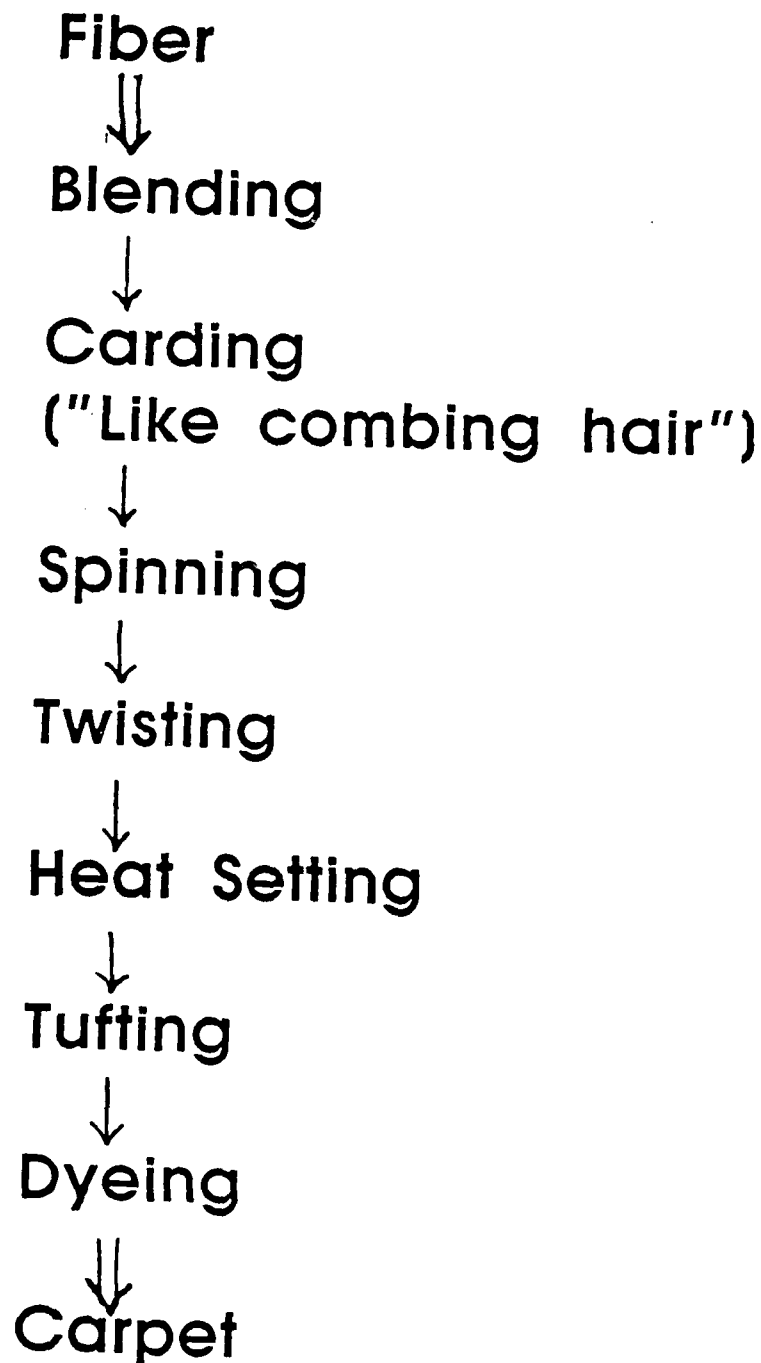


5

Customers' Processing

Example: Bales of Fibers (8") used
for Carpets

Steps:



Special Problems/Blessings in Chemical Processing

- **Multiple Processes**
- **Variation Smoothing thru Blending**
- **Upsets - Large Amounts of Product Loss**
- **Offgrade Product - Possibly Hazardous Waste Problem**
- **Impurities can Build-up in Process (Non-Steady State)**
- **Within Spec is Not Enough -- On-Target is Better and Often Necessary (e.g., Carpet Yarn)**

Steps in Quality Improvement

1. Characterization ("Landscaping")

- Flow Diagrams
- Historical Data Analysis
- Pareto Analysis
- Correlation (x-y Plots)
- Process Capability
- Measurement Capability

2. Plant Trials (Cause and Effect)

- Fishbone Diagram
- Variance Components
- On-Line/Off-Line Experimentation

3. Modeling (Prediction)

- Process Simulation Models

**Example: Handling Uncertainty in Engineering
Design and Quality Improvement**

- 1. Redesign or Reformulation of Product**
- 2. Redesign of process**
- 3. Design of Better Inventory Control
System**
- 4. Design of Better Customer Blending
System**

Benefits:

- 1) More 1st Quality Product**
- 2) Reduced Variation**
- 3) Customer Satisfaction**

CASE STUDY #1

Objective of Case Study

- Improve Product Uniformity on
Nylon Fiber

Benefits

- Customer Satisfaction
- Reduced Claim Costs
- Increased "Net First Quality" Product

Approach

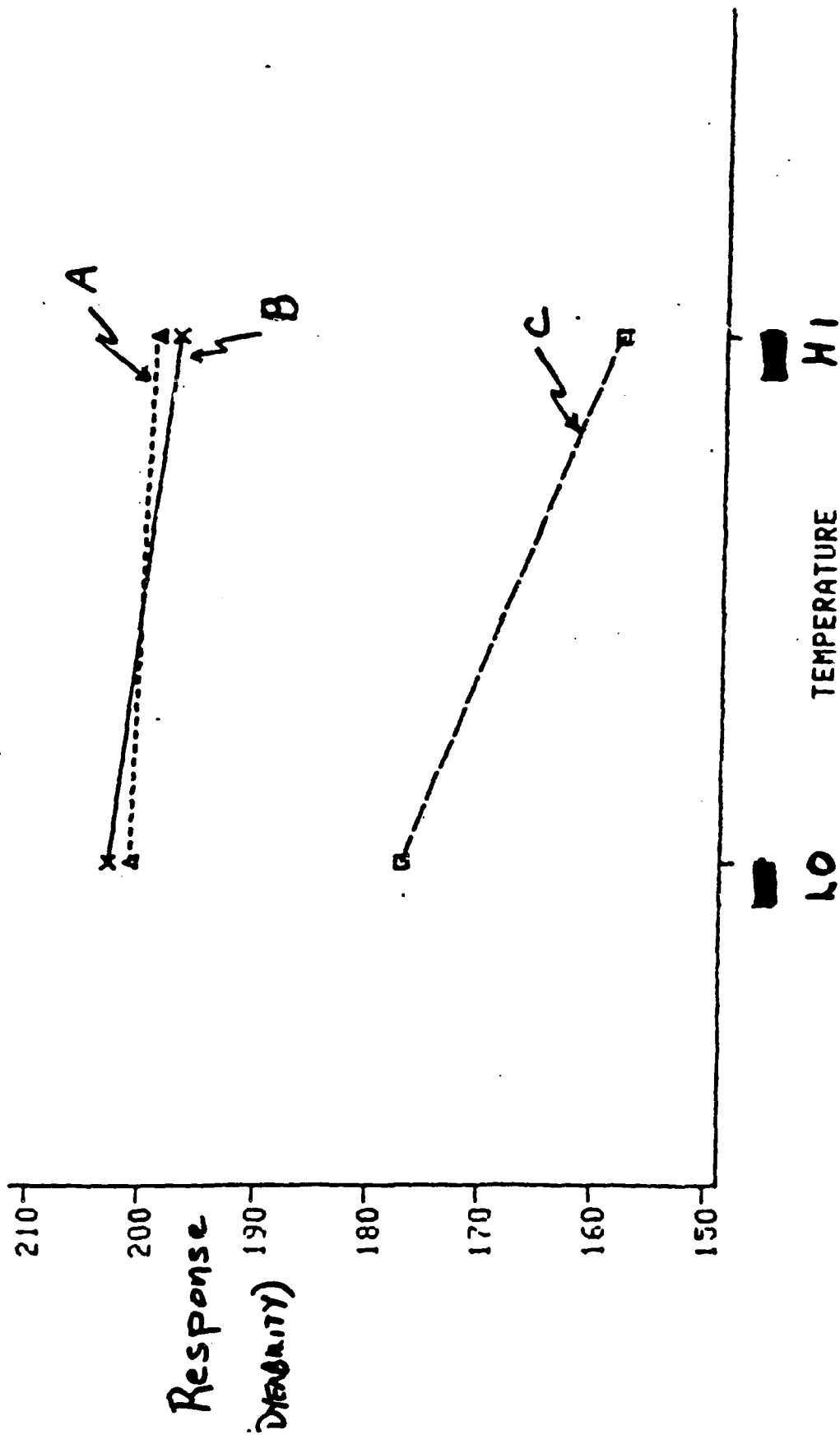
Historical Data Review

- Expert Opinion
- Literature Search
- "Cause-and-Effect" Diagrams
 - 125 Variables
- "Pareto" Principles
 - e.g., - 80% of off-quality material was from product "C"

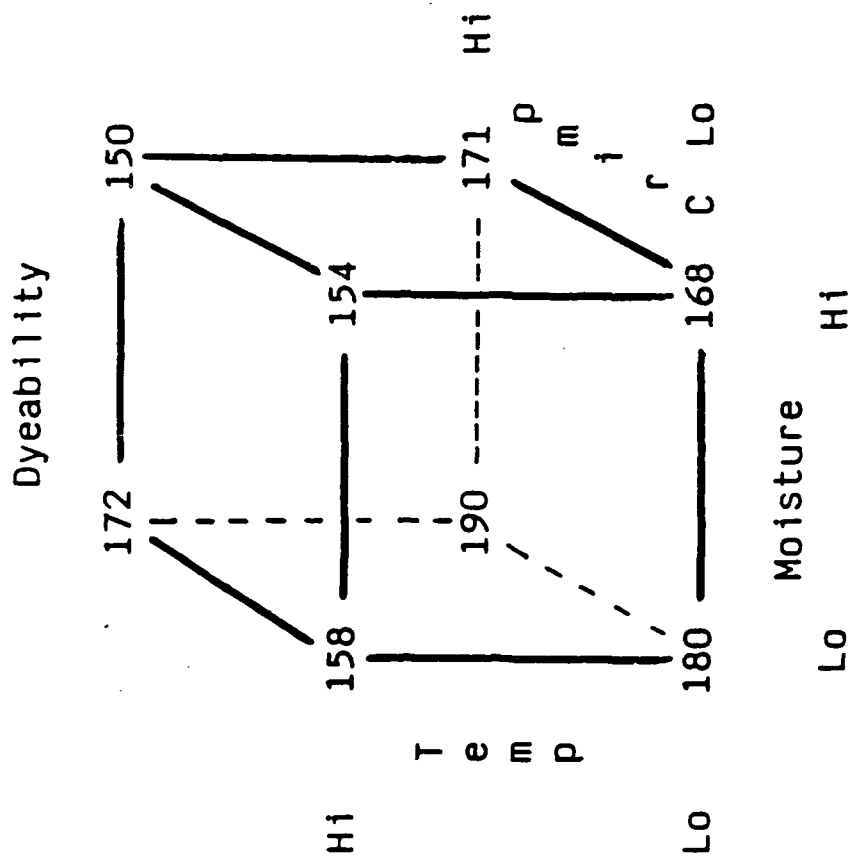
Statistically Designed Experiments

46 of 125 Variables Studied

Effect of Temperature



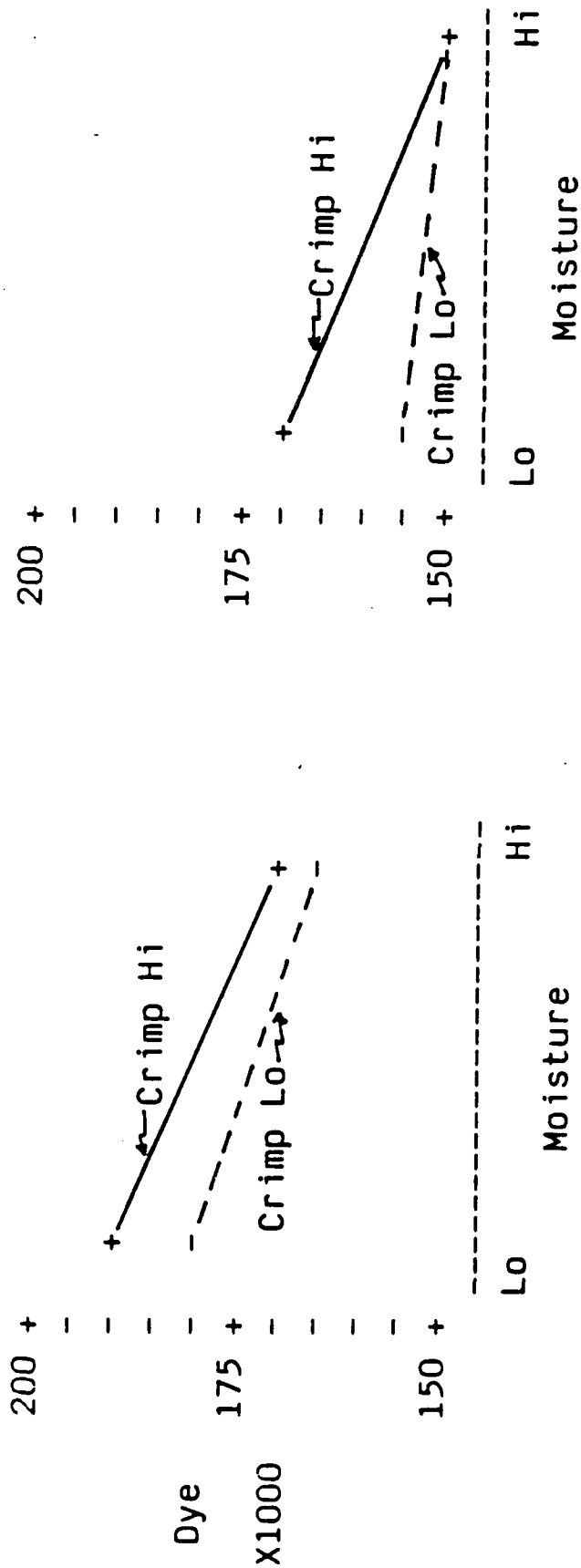
EXAMPLE DATA



Changes in Dyeability
Due to Variables and Interactions

Moisture (M)	-14	±	7.6	4.0
Temperature (T)	-19		"	4.0
Crimp (C)	6		"	7.6
MXC	1		"	
MXC	-6		"	
TXC	-1		"	
MXCXC	-3		"	

Dyeability



Temperature= Lo

Temperature= Hi

RESULTS

1) Product "C" Redesigned to be Robust
to Customers' Processes

2) Related Quality Costs Reduced by
Factor of 10

3) Net First Quality Product

1984	86%
1985	94%
1987	97%

4) High Customer Satisfaction

**"The real job of any business is
creating value"**

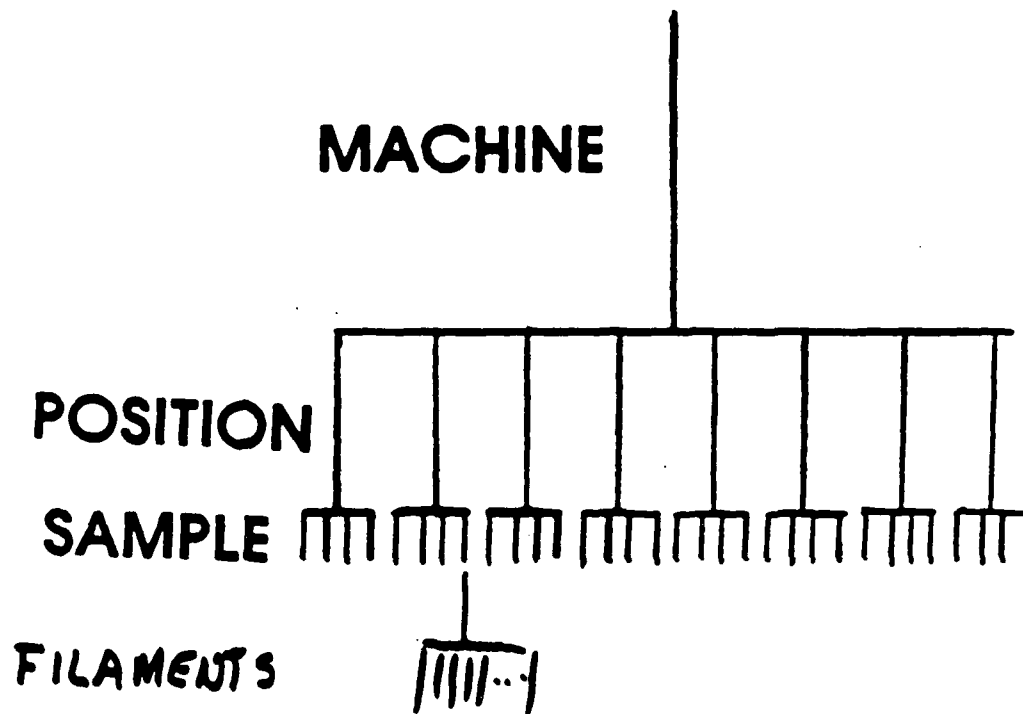
**Alan Belzer
President, Engineered Materials
Allied-Signal Inc.**

(Ref: Chemical Week, March 11, 1987)

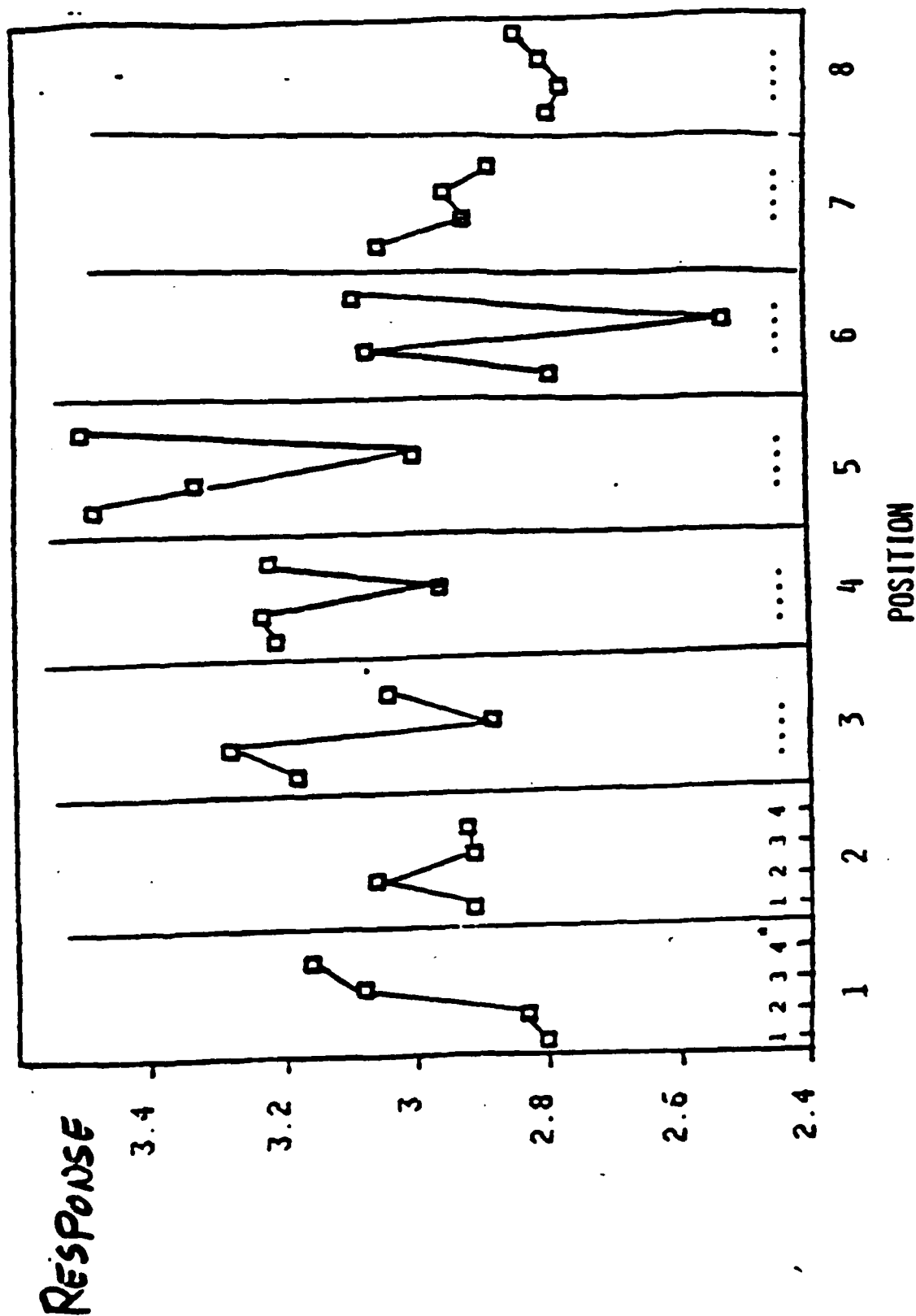
VARIANCE COMPONENT DESIGN FOR PROCESS

- SAMPLE-TO-SAMPLE (TIME)
- POSITION-TO-POSITION
- FILAMENT-TO-FILAMENT

FILAMENT-TO-FILAMENT AUTOMATICALLY
ESTIMATED BY INSTRUMENT. ABOUT 50
FILAMENTS PER SAMPLE.



STATISTICAL DESIGN RESULTS



*Sample Sequence.
Samples taken every 12 hours
over 2 days.

BREAKDOWN OF VARIATION SOURCES

		STD. DEV.
POSITION-TO-POSITION	σ_p	.15
SAMPLE-TO-SAMPLE	σ_s	.09
POSITION X SAMPLE	σ_{ps}	.14
FIL-TO-FIL	σ_f	.38

STANDARD DEVIATION FOR
FUTURE PRODUCTION SAMPLE=

$$\sigma = \sqrt{\left(\frac{p-1}{p}\right) \frac{\sigma_p^2}{N} + \sigma_s^2 + \left(\frac{p-1}{p}\right) \frac{\sigma_{ps}^2}{N} + \frac{\sigma_f^2}{N}}$$

WHERE p = # OF POSITIONS
 N = # OF FILAMENTS

RESULT:

PROCESS REDESIGN

& CHANGES

COMPUTER MODELING & SIMULATION

Important when:

- Experimentation is expensive
- Resources (e.g., capital & people) are scarce
- Testing factors outside of our control
- Predicting product performance at customers
- Obtaining management support

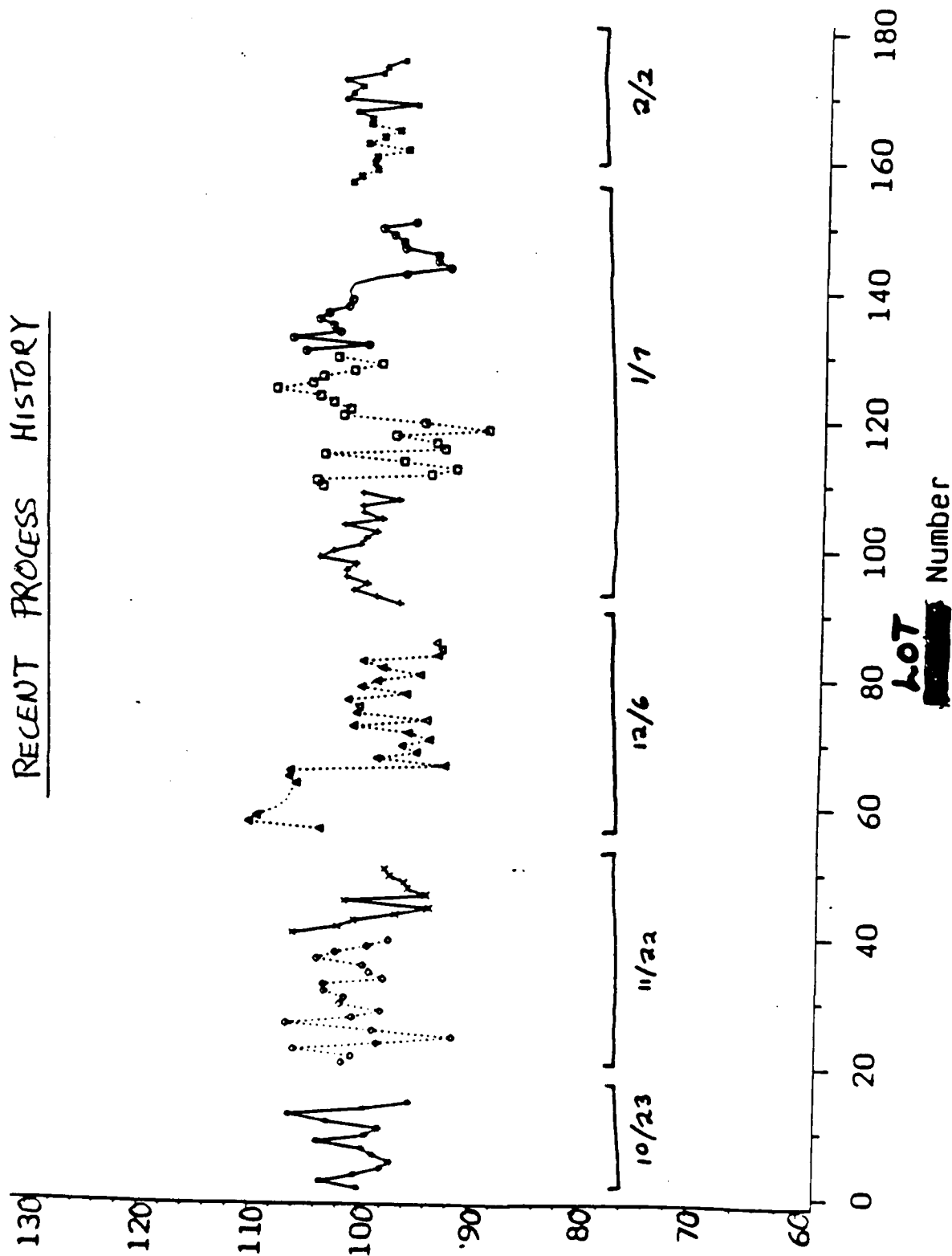
Illustration - Computer Simulation

Box

**"All models are wrong,
some are useful"**

MEASURING IMPACT OF INCOMING VARIATION

RECENT PROCESS HISTORY



Random Variation and Autocorrelation

- Fiber produced in batches called "creels"; usually 8 bales
- Creel averages may wander over time
- Variance component model for armful a of bale b of creel c :
(package)

$$y_{cba} = \mu + C_c + B_{b:c} + A_{a:bc}$$

$$B_{b:c} \sim i.i.d. N(0, \sigma_b^2)$$

$$A_{a:bc} \sim i.i.d. N(0, \sigma_a^2)$$

$$C_c \sim \text{normal AR}(1); \phi, \sigma_c^2 = \text{Var}(C_c)$$

Range of poor to good stat. control

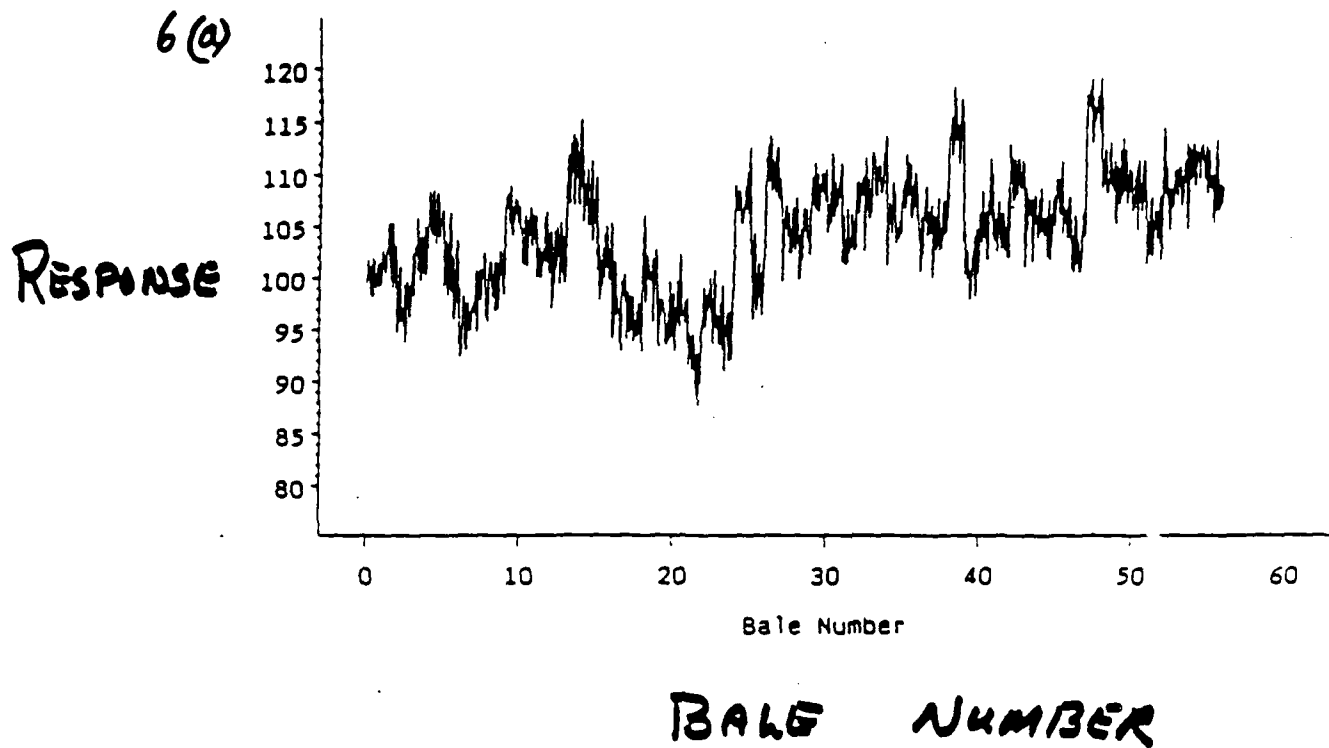
26

Example : (approx. values from real data)

Parameter	Value
μ	100
σ_c	8
σ_b	3
σ_a	2
ϕ	.75

Figure 6: Plant A: Simulated variation in incoming fiber.

Plant A: Autoregressive Model for Initial Material



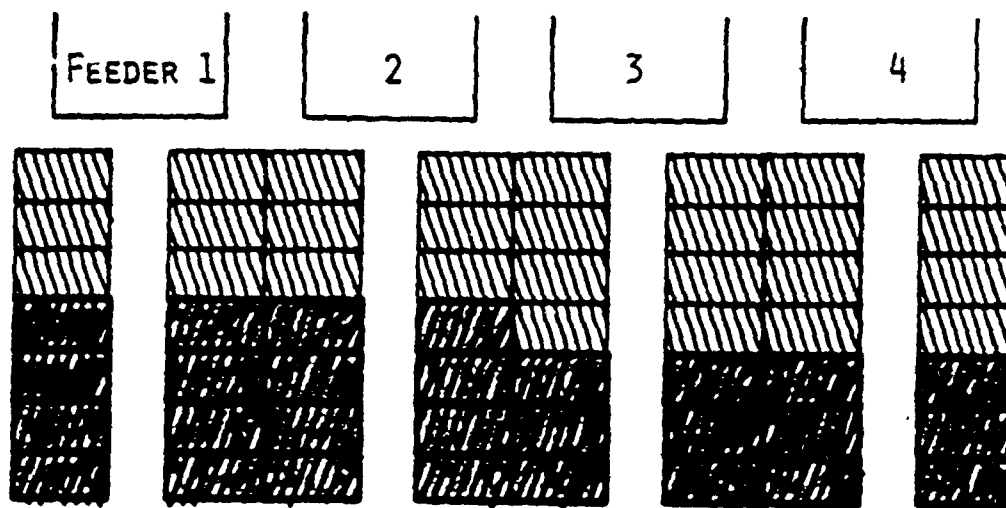
10029 8125

Variables Affecting Uniformity

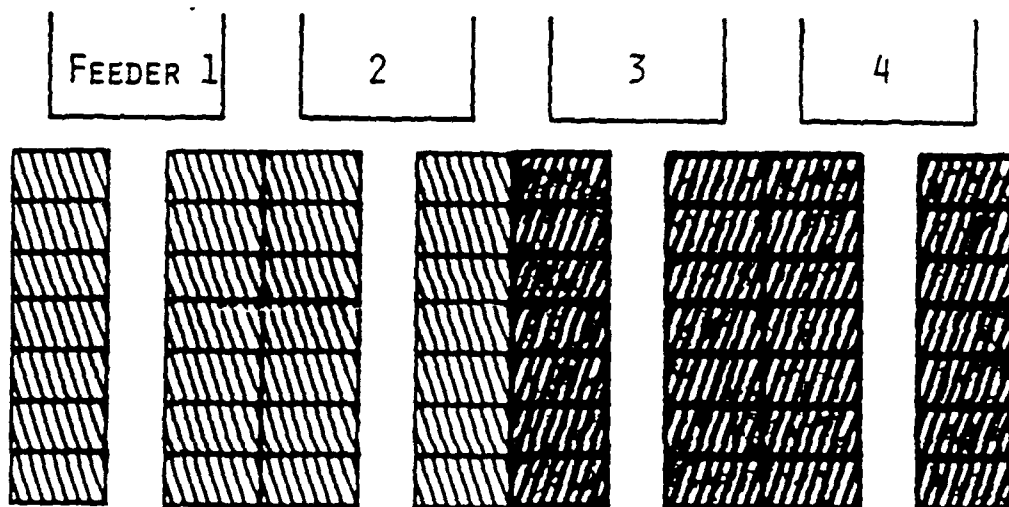
- Type of initial non-uniformity
- Inventory/shipping management
- Blending pattern
 - lot size
 - number of feeders
 - laydown pattern
 - armful size
 - cross-blender?
 - ...
- Carding pattern



CASE 1. LAYDOWN BY ROWS FROM TRUCK



CASES 2&3 LAYDOWN BY FILES FROM TRUCK ^{OR COLUMNS}

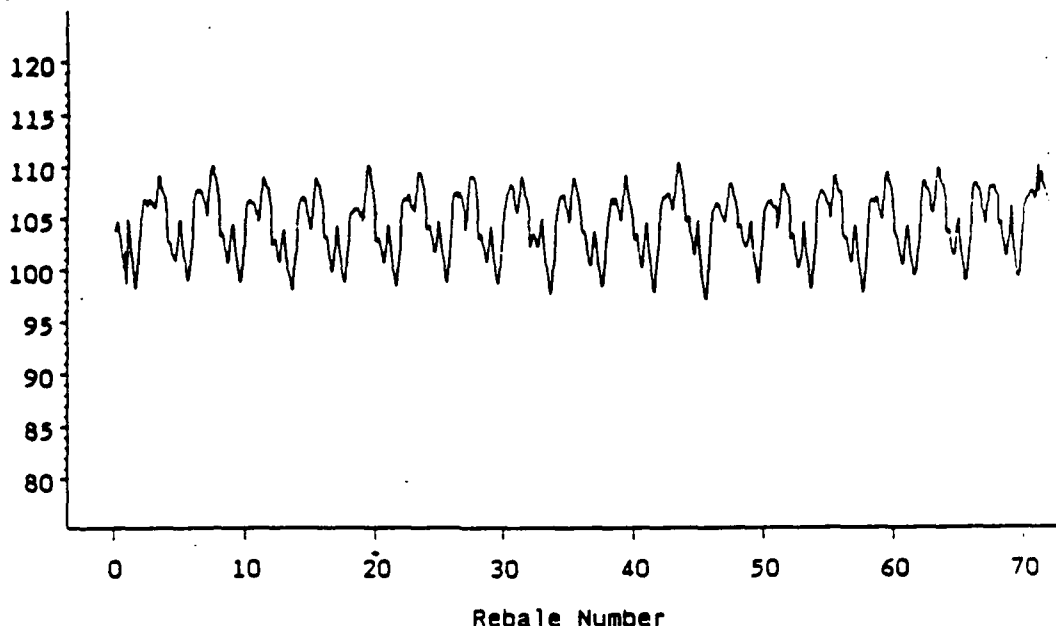


6 (b)

Laydown by rows
After First Blending Pass

1ST PASS

RESPONSE



BALE #

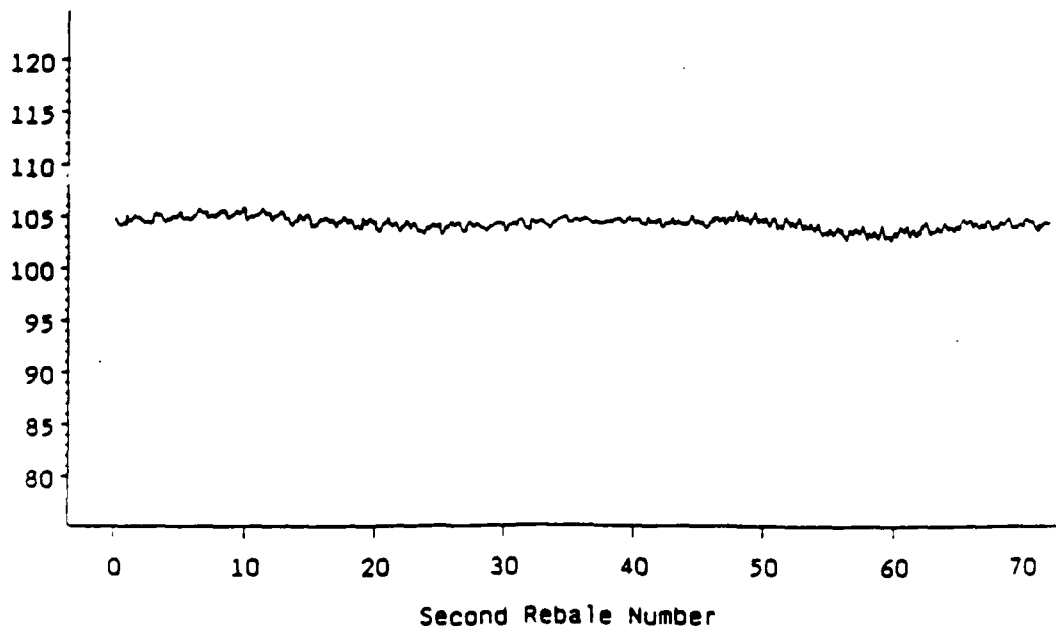
10029 6129

6 (c)

Laydown by rows
After Second Blending Pass

2ND PASS

RESPONSE



BALE #

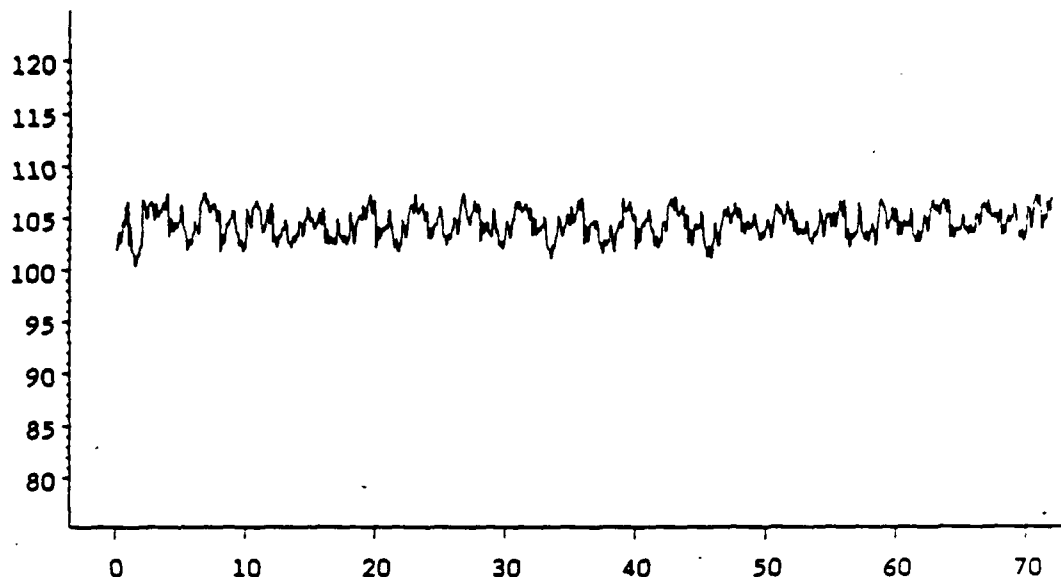
10029 6132

Laydown by Columns
After First Blending Pass

1ST PASS

6 (d)

RESPONSE



Rebale Number

BALE #

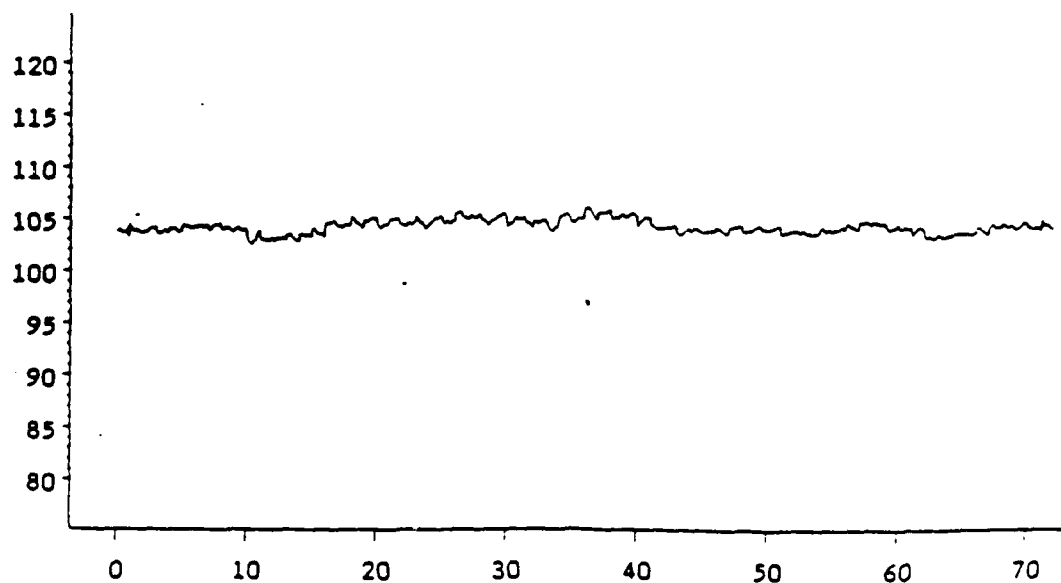
10029 6143

Laydown by Columns
After Second Blending Pass

2ND PASS

6(e)

RESPONSE



Second Rebale Number

BALE #

10029 6145

Results from Simulated Random Variation

First pass :

- laydown by columns much better than by rows

Second pass :

- laydown by rows vs. columns gives similar results
- uniformity improved over first pass

**RESULT: CERTAIN CUSTOMERS IMPLEMENTING
REVISED BLENDING EQUIPMENT
DESIGN CONFIGURATIONS**

WRAP-UP

- o Quality Improvement via Statistical Methods that Treat Uncertainty
- o Chemical Product and Process Design/Redesign/Improvement via Understanding the Impact of Uncertainty on Quality

In the Spirit of Bill Hunter's Earlier Quote

Quality will improve as we
match our march to
nature's

Session III.

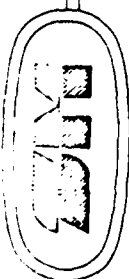
Statistics in Design
for Flexible Automation

Date: March 17th, 1988

Time: 7:32 pm

Filename: 7364-27.L43

Remark: SHEFFIELD (JOB# 8316) - DISK D



AUTOMATED DIMENSIONAL INSPECTION SYSTEMS

by
John A. Bosch
Stephen L. Fix
Sheffield Measurement

May 11, 1988

NBS / GVV Conference on
Uncertainty in Engineering Design
Gaithersburg, Maryland



As Global Competition Intensifies...

- Higher Quality
- Increased Productivity

Creates a Need for...

Automatic Dimensional Inspection Systems



Major Inspection Trends...

- On Factory Floor
- By Production Personnel
- For Process Control

To Provide...

Quality Parts - On Time



Flexible Inspection Systems Offer...

- Ability to Measure a Variety of Parts
- Shorter Lead Time for Changes or New Parts
- Lower Capital Expenditure
- Absolute Measurements for Trend Data
- Automation Flexibility
- Integration with Factory Host - CAD / CAM

For...

Better Process Monitoring and Control



The Manufacturing Engineer Has Many Options...

- Dedicated Gaging
- Vision Systems
- On Machine Gaging
- First Piece Inspection (Off-Line)

With the Most Significant Trend Being...

Flexible Inspection Systems



FIS Brings New Requirements...

- Environmental (Temp., Vib., Atmospheric)
- 3-D Gage Surveillance
- Up-Time
- Parts Handling
- Communication Interfaces

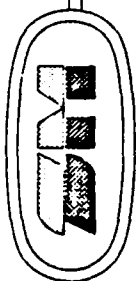
Creating the Need for...

Timely Inspection Planning



Measurement Equipment Accuracy

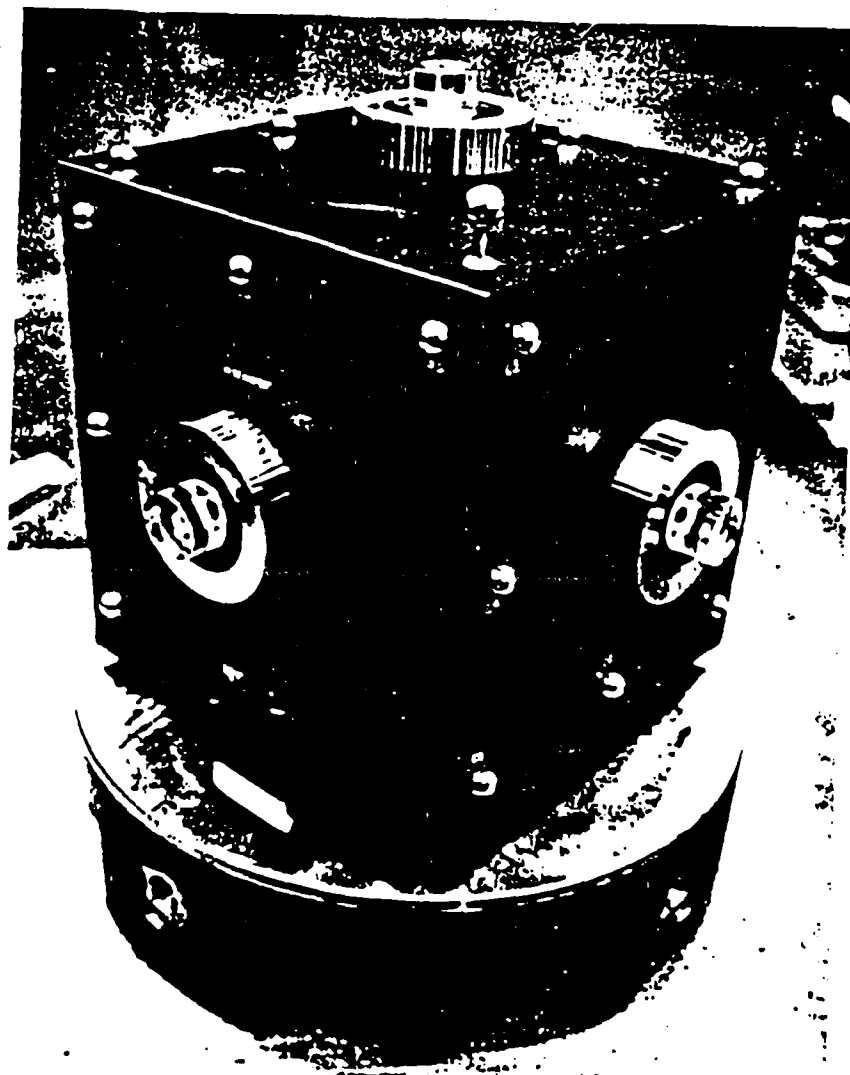
- Must Be Four to Ten Times More Accurate Than Tolerance Specified
- Compounding Features Magnify Individual Measuring Accuracy Requirements
- Repeatability and Reproducibility Requirements Force Better Control of the Measurement Process



Measurement Process Uncertainty

- Traceability Is Critical
- Environmental Factors Must Be Considered
- Geometric Algorithms Need Validation
- Part Measuring Accuracy is Fundamental Criterion

Requires Periodic Verification





Process Control Involves...

- Knowledge of the System
- Dynamic Effects
- Environment Factors
- Traceability of Standard

To Establish Level of
Uncertainty in the Process

SDA

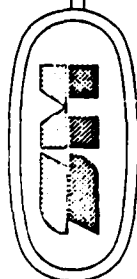
Real Time Statistics Shows

- Actual Data of Part and Process
- Determine Capability of Process
- Show Data Trending or Stratification

With a Goal of...

No Bad Parts

//



Exception Report

Part Program: Housing 3

Revision: 01.02

Number of Exceptions: 5

Page: 1

Date: 02/15/88

Time: 10:39

Seq	Dim	Feat-Name	Ftype	Otol	%BPT	XB-CL	TX	Cp	Cpk
00001	Y	FEAT001	PNT					.93	
00002	DI	HOLE001	I.D.		-91.3%		1-		-.65
00003	DS	DIST002	DIST	.0002	120.3%	0.0001			
00004	Z	PROCESS3	SPHR		93.1%		2+	1.12	.98
00005	X	PLANE02	PLNE				ST		

*** end of exceptions ***

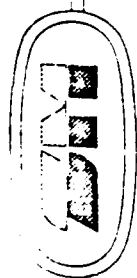
Note: Only 1 Feature, Sequence 0003, Exceeds Specification Limits.

Date: March 21st, 1988

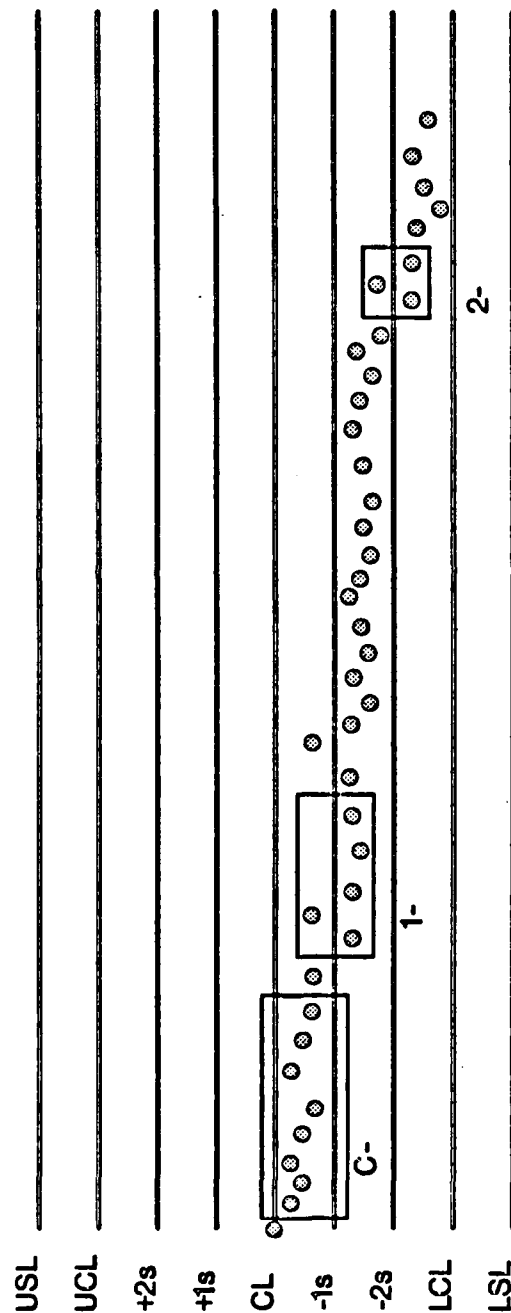
Time: 11:07 am

Filename: 7364-25.L43

Remark: SHEFFIELD (JOB# 8316) - DISK D

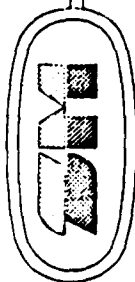


Drill Bit Wear Hole Diameter

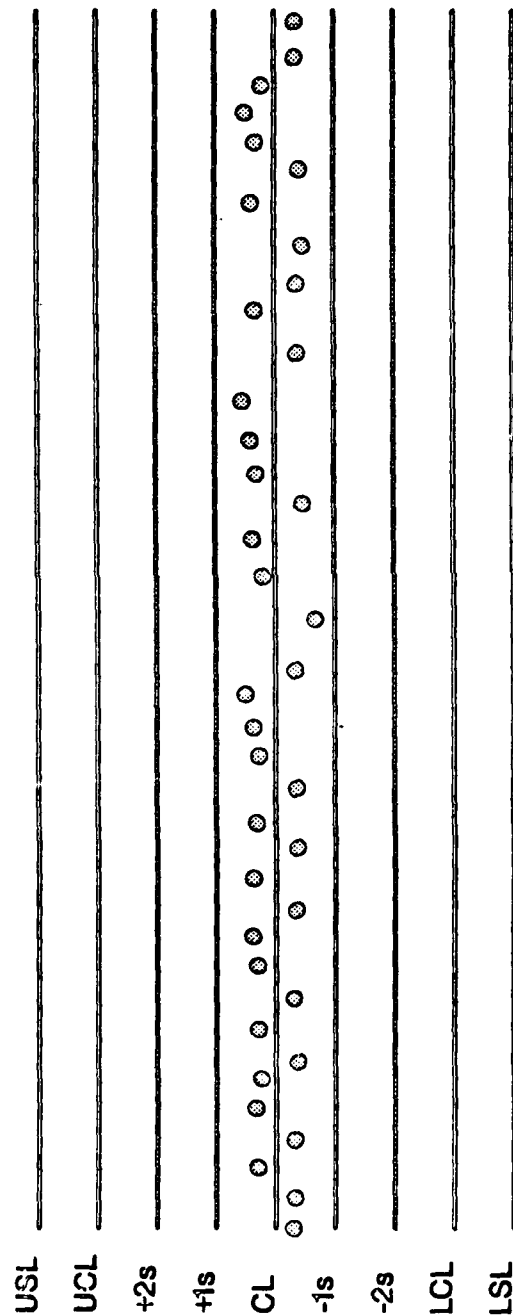


C-	1-	2-
8 Consecutive	4 of Last 5	2 of Last 3
Subgroups Below	Subgroups Are	Subgroups Are
Center Line	Below -1 Sigma	Below -2 Sigma

Remark: SHIEFFIELD (JOB# 8316) - DISK D



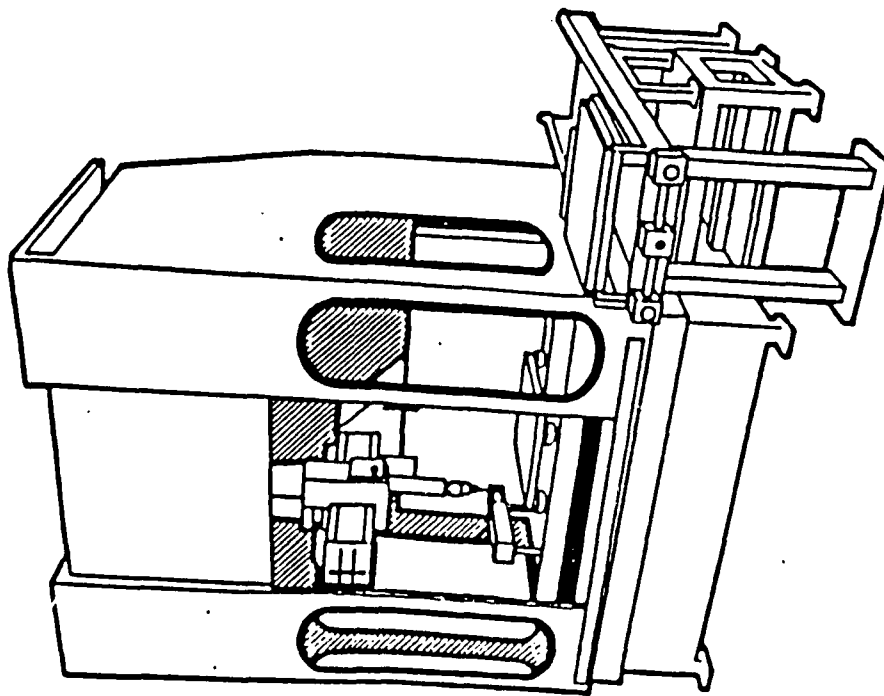
Stratification



Causes: Non-Random Sampling
Pre-Screening
Improper Control Limits



STANDARD FIS SYSTEMS ARE BEING OFFERED



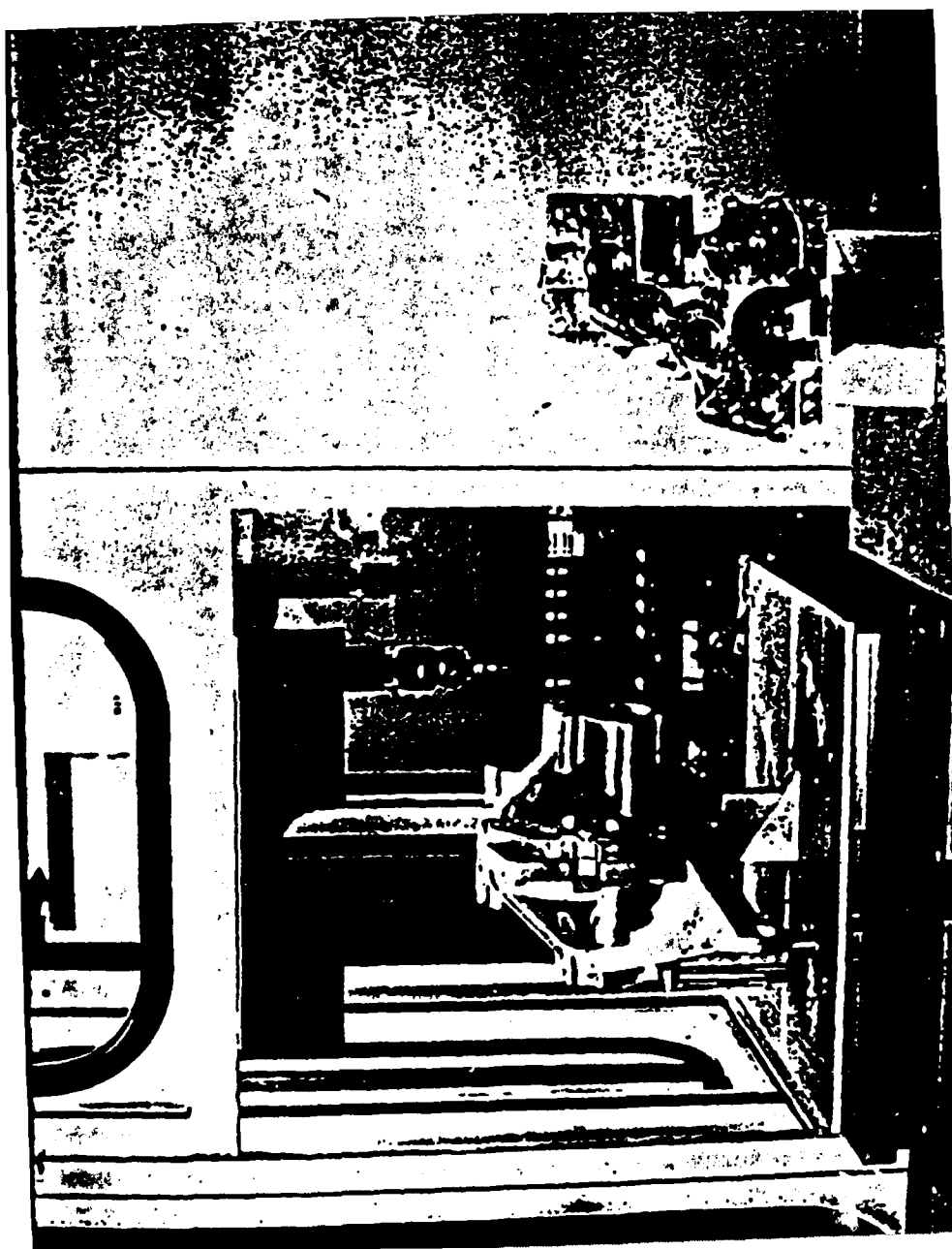
With Automatic...

- Operation
- Probe Changing
- Temperature Compensation
- Part Handling
- Part Recognition
- Statistics
- Probe Calibration

In an Enclosure for Use on
Shop Floor



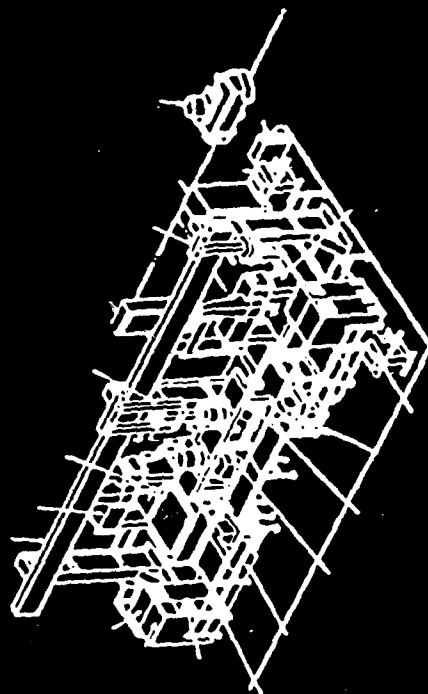
16



1/2/91

Individual FLS Cells Offer...

- Palletized Part Loading
- Job Buffering and Scheduling
- Cell Management
- Factory Host Communication Link
- Relational Data Base Management
- Extensive SPC Functions
- CMM CAD INTERFACE-DMIS







Distributed Computer Architecture Provides...

Simultaneous Operation

Data Collection

Geometric Calculations

Automatic Machine Operation

Probe Articulation

Program Development

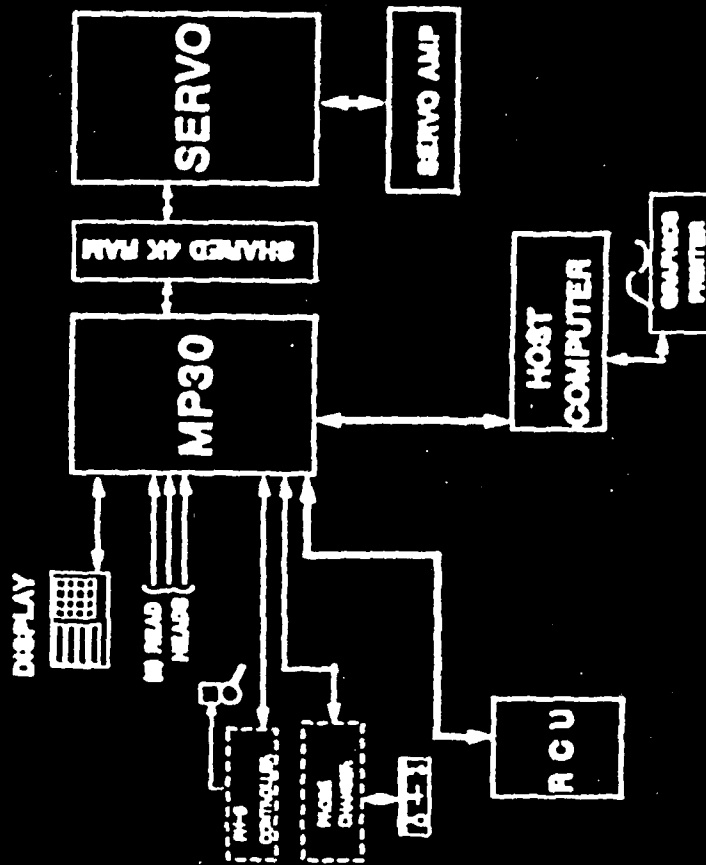
Report Generation

For

Fast Throughput

NAV

CORDAX APOLLO Diagram: MP-35 System

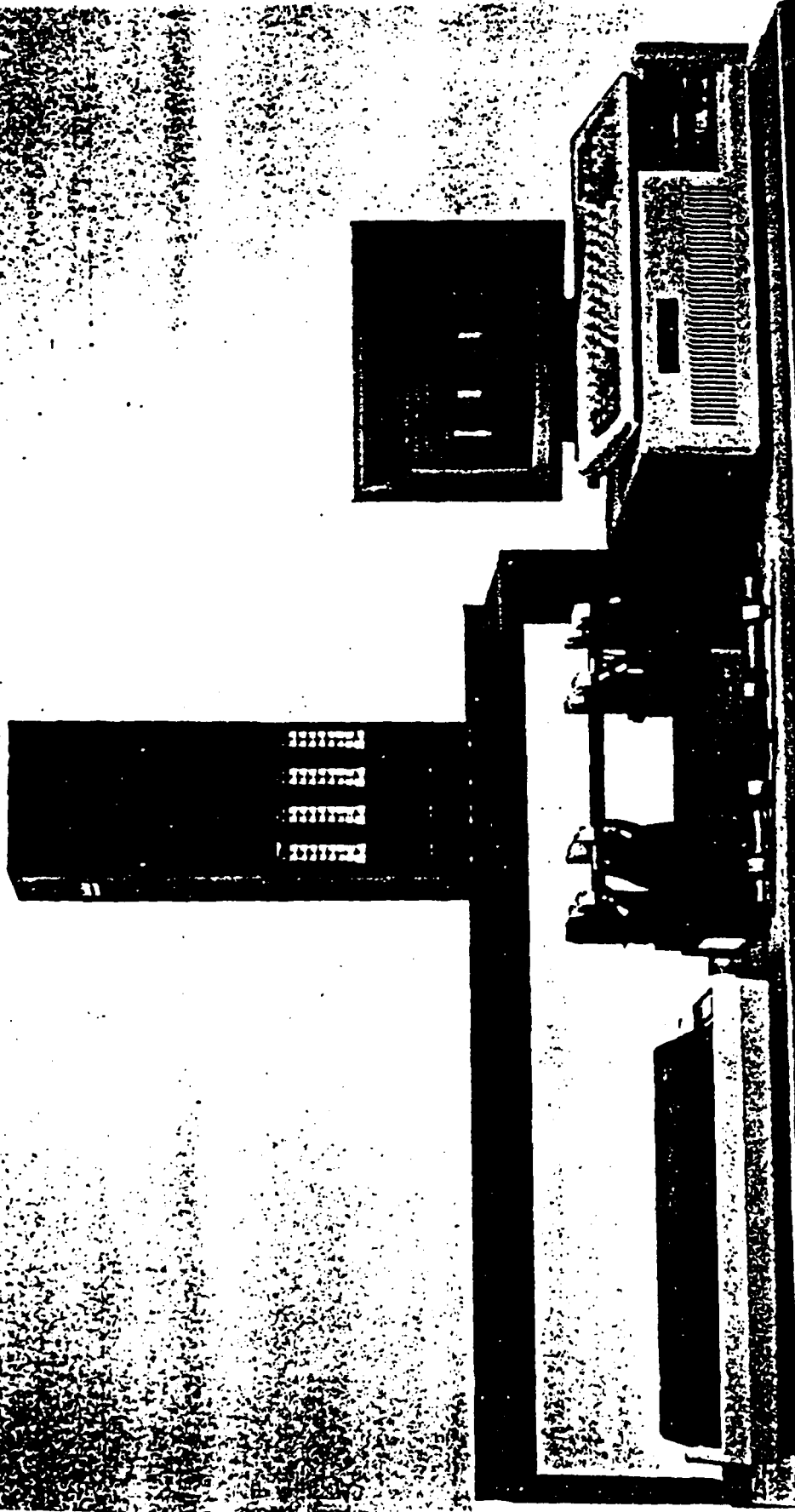


EV

Dedicated Gages Still Provide Advantages...

- High Speed
- 100% Inspection
- Precision Measurements
- Near Real-Time Control





2511 Sheffield



Vision Gaging Systems Are Worthy of Consideration for...

- Off - Line High Precision Parts
Mainly 2-D
- In - Line Inspection Tasks



CONTROLLED MACHINING BY CLOSING THE LOOP THROUGH AN FIS REMAINS ELUSIVE...

Major Obstacles Exist...

- Timely Data
- Extraction of Data of Interest
- Determination of Corrective Action

For Now...

Human Intervention is Required -
But Made Easier by FIS



Uncertainty In Process Control Is...

- Analyzed by Process Variance Methods
- Reduced by Statistical Methods
- Controlled by Near Real - Time Measurement
- Achieved by Employee Involvement

To Fully Realize the Design



Quality Involves Teamwork Through...

- Management Commitment
- Employee Involvement

With a Key Element Being...

Automatic
Dimensional Inspection Systems

What is a Tolerance? The Problem of Methods Divergence in Flexible Automation

**Dr. Theodore H. Hopp
National Bureau of Standards**

**Conference on Uncertainty
in Engineering Design**

May 10-11, 1988

Methods Divergence is a Problem

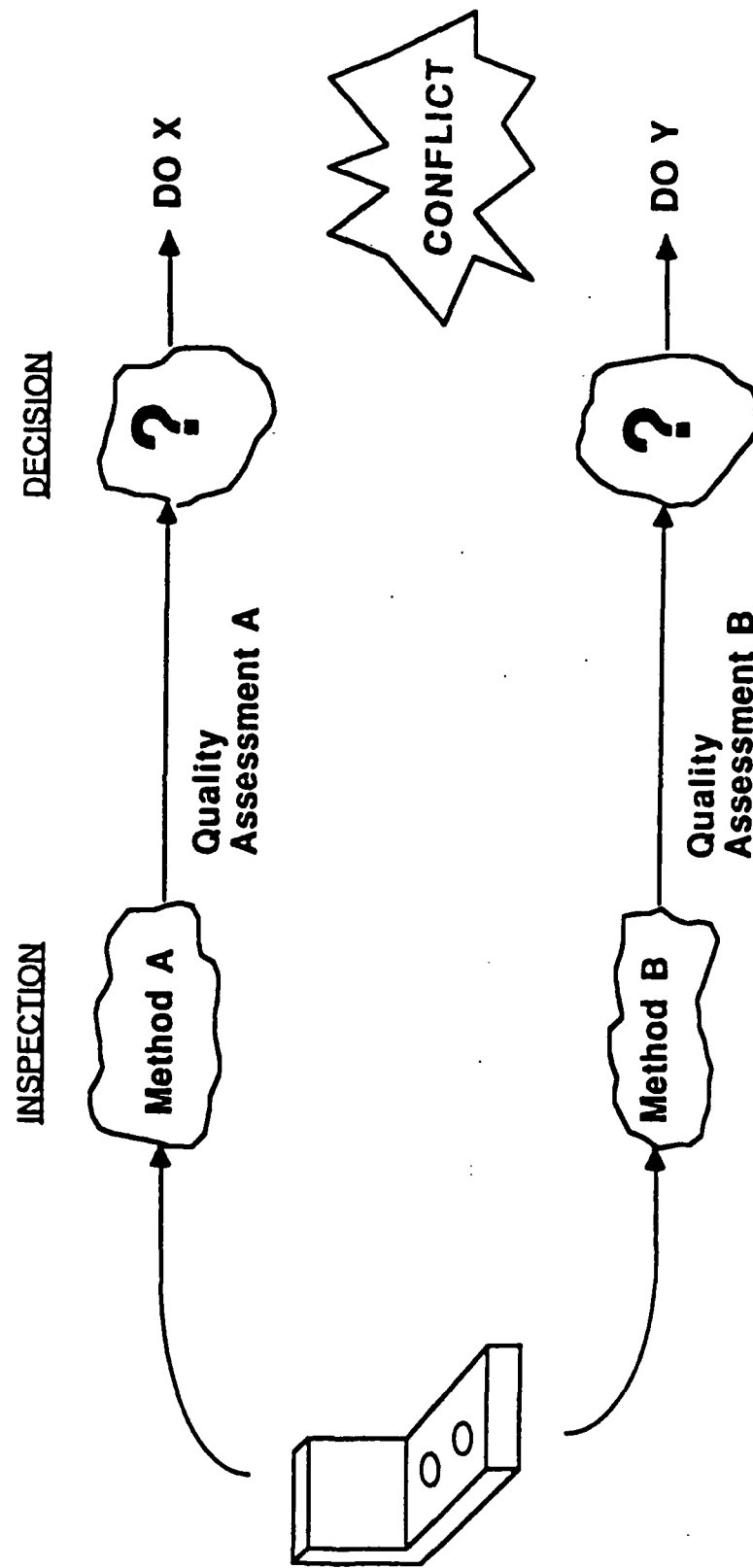
Item: Rock Island Arsenal institutes a policy of replacing functional gages with coordinate measuring machines.

Reason: High cost and inflexibility of functional gaging.

Item: DCAS shuts down the production line of a major defense contractor for several days.

Reason: Lack of control over the quality of data analysis software on coordinate measuring machines.

Definition of Methods Divergence



What is Needed:

**A Rational Basis for Choosing Among
Alternative Technologies for
Quality Assurance**

Key Concepts

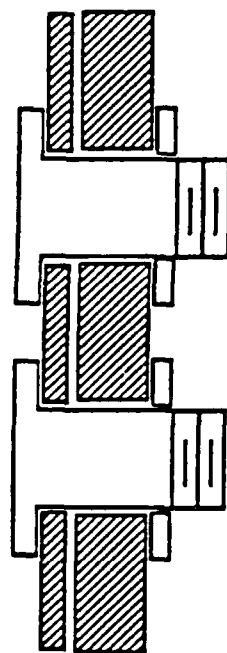
Design Intent

Dimensioning & Tolerancing (ANSI Y14.5)

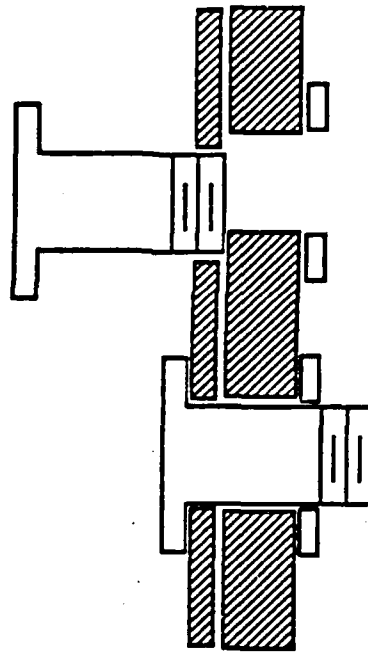
Inspection Systems

- Functional Gaging**
- Sample Measurement**

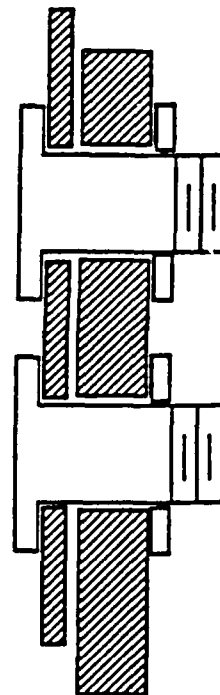
Design Intent: Interchangeable Assembly



Good



Misaligned
Holes



Shifted
Plates

Decoupling Techniques

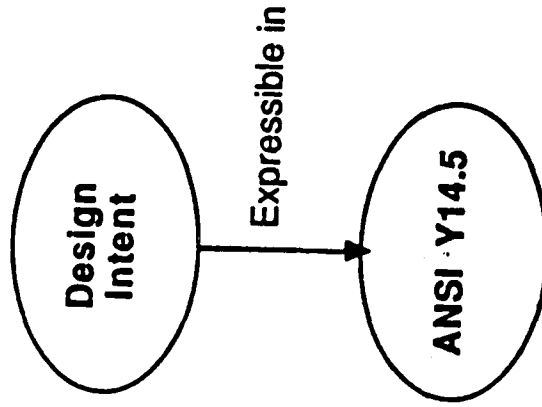
Standard Components

Tolerance Allocation Among Components

Conditional Tolerancing

Quality Assurance Folklore, Scene 1

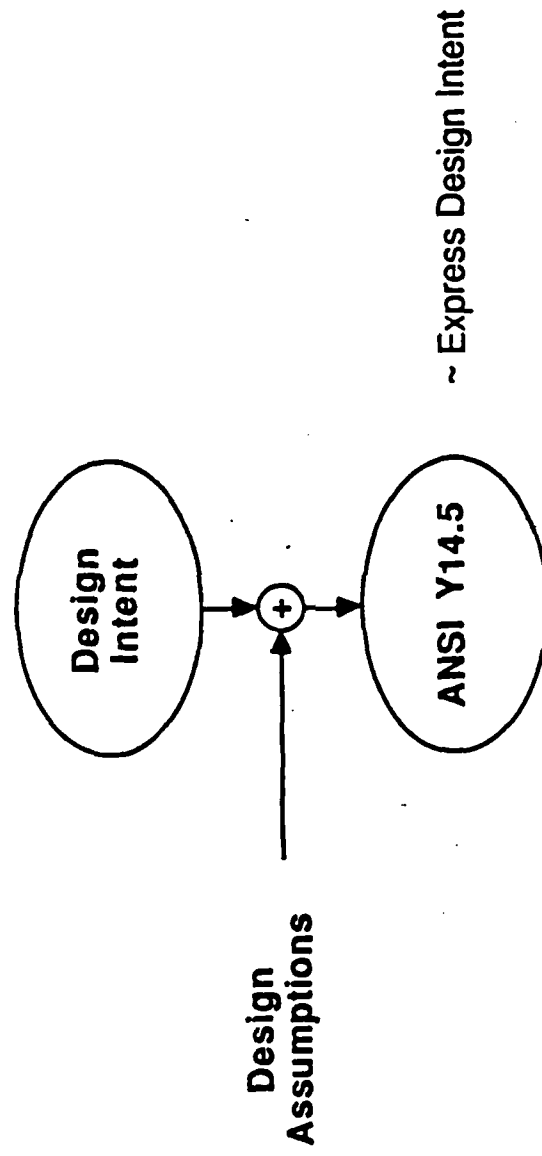
The Common View



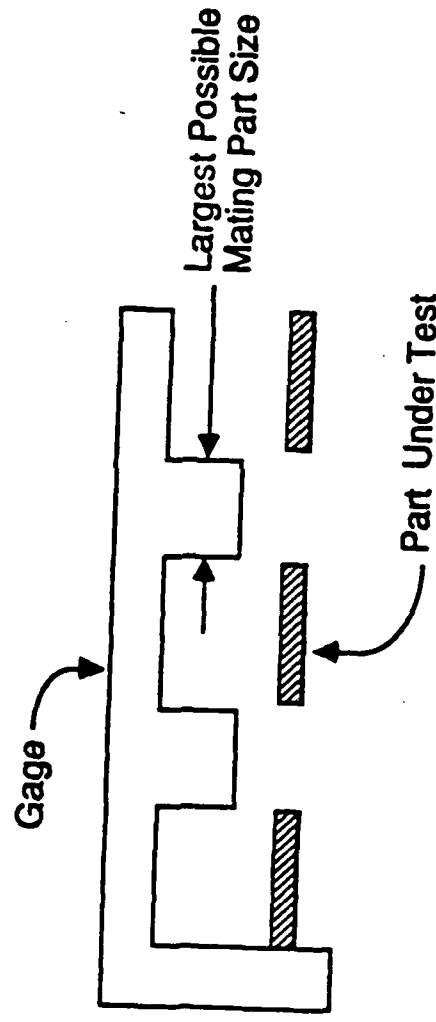
The Fallacy

ANSI Y14.5 requires
decoupling assumptions
to express design intent

Quality Assurance Reality, Scene 1

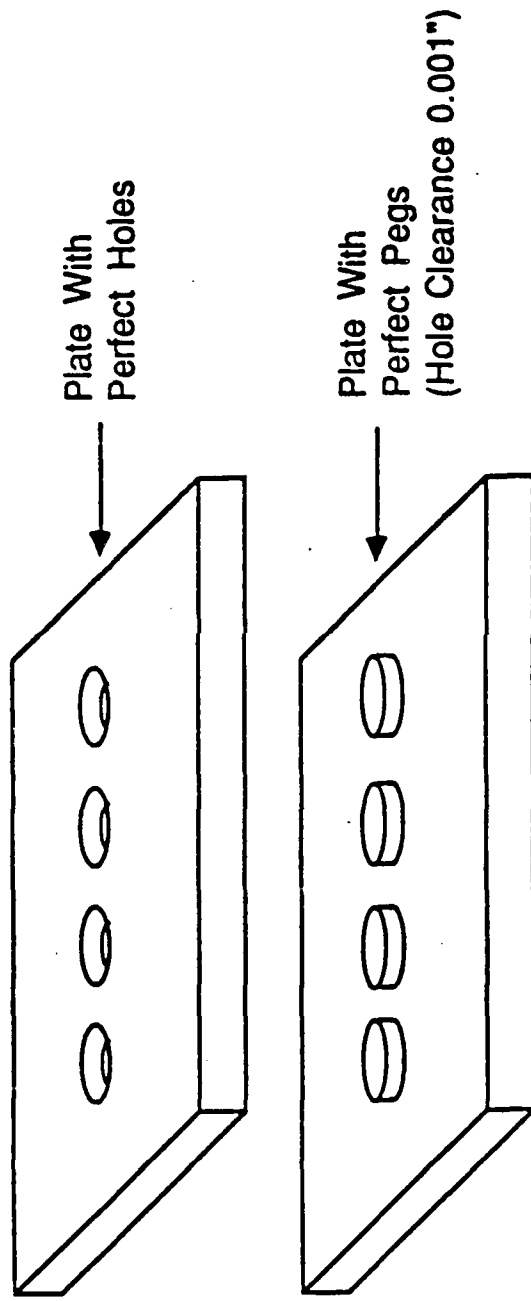


Functional Gaging Systems: Simulation of Worst-Case Mating Conditions



Errors in Functional Gage Systems

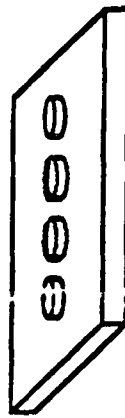
(With Apologies to H. David)



Experiment: Add position errors to two pegs. Investigate $Pr(\text{fit})$ as a function of which pegs get the error.

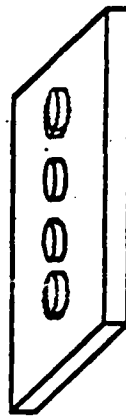
Errors in Functional Gage Systems

<u>Simulated Source of Error</u>	<u>Probability of Fit</u>
----------------------------------	---------------------------



Central Pegs

0.292



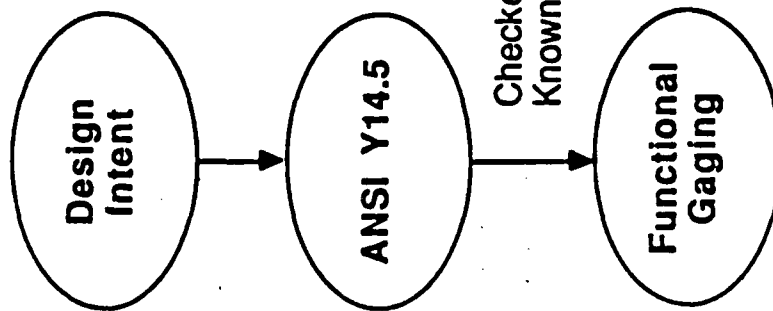
Outer Pegs

0.454

Quality Assurance Folklore, Scene 2

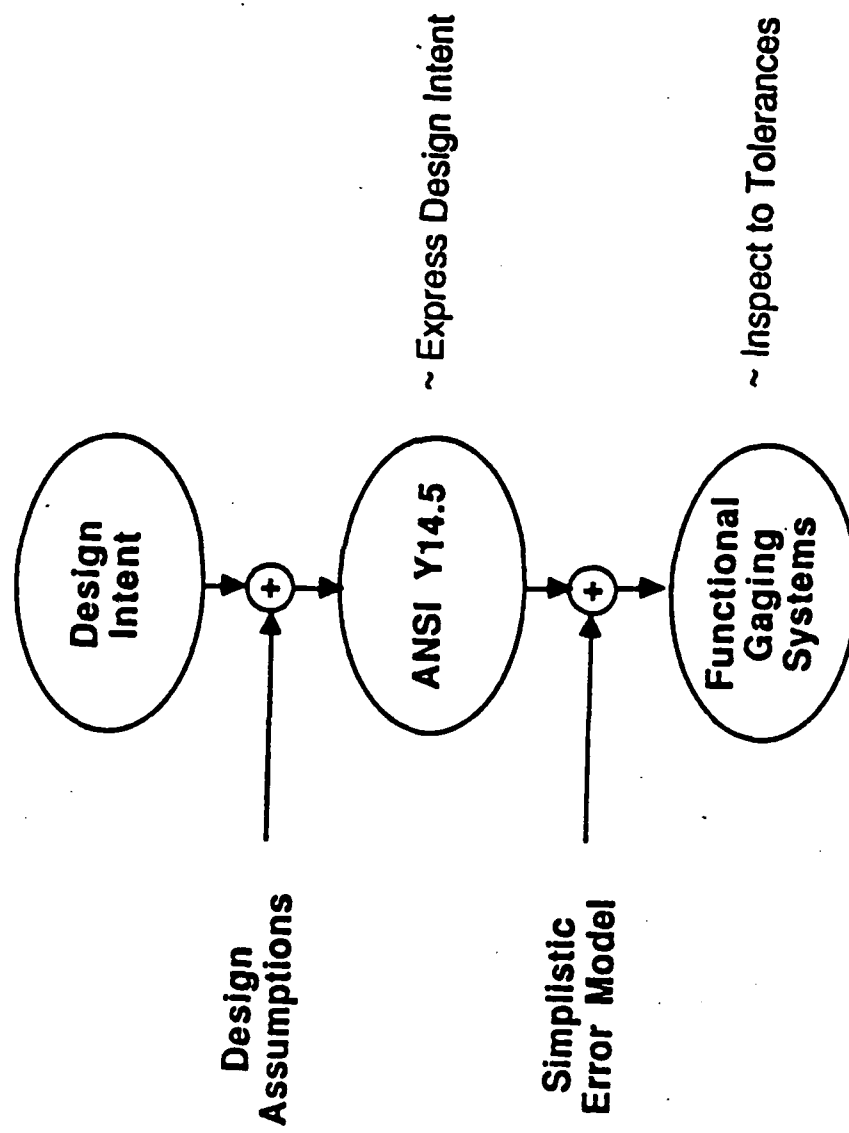
The Common View

The Fallacy

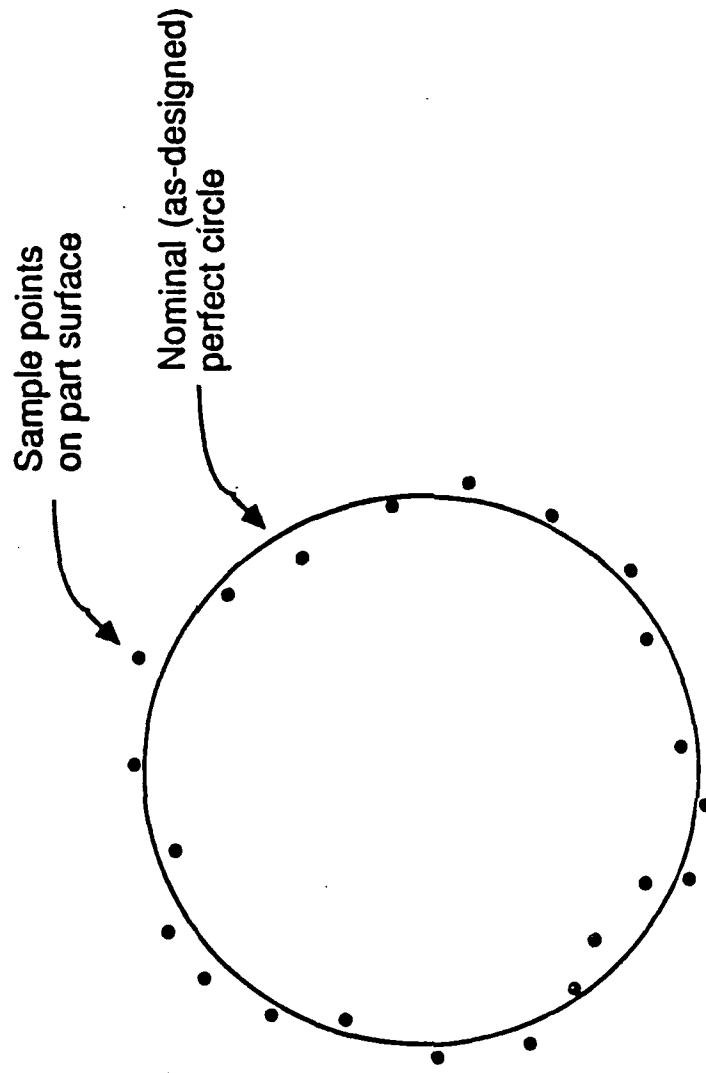


The effects of "critical geometry" on gage tolerance are ignored

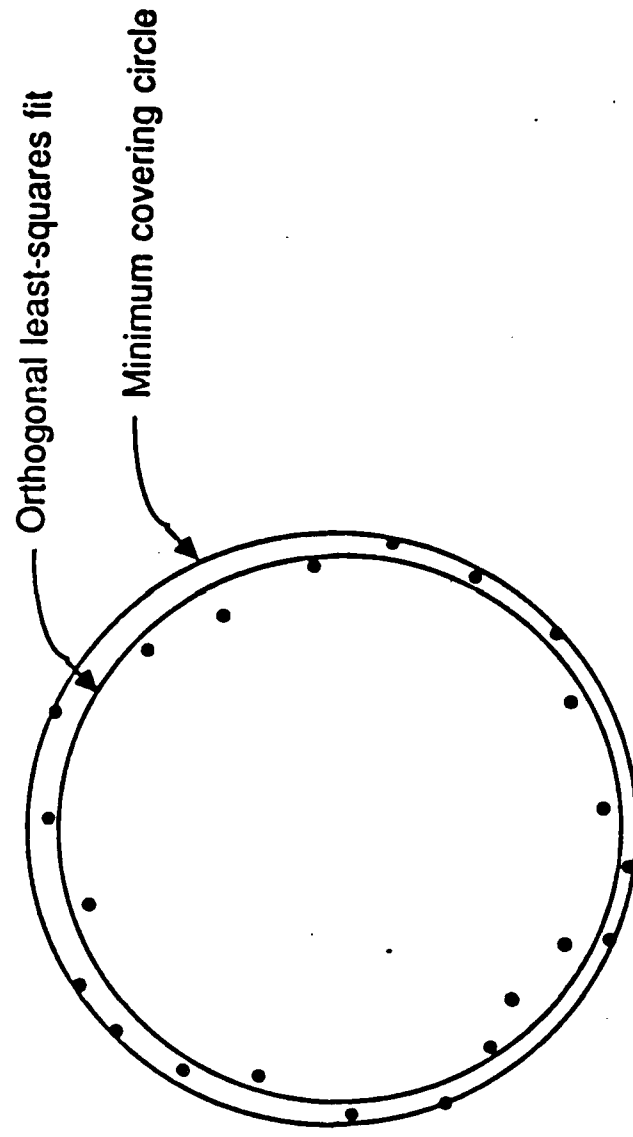
Quality Assurance Reality, Scene 2



Sample Measurement Systems: Analysis of Point Data

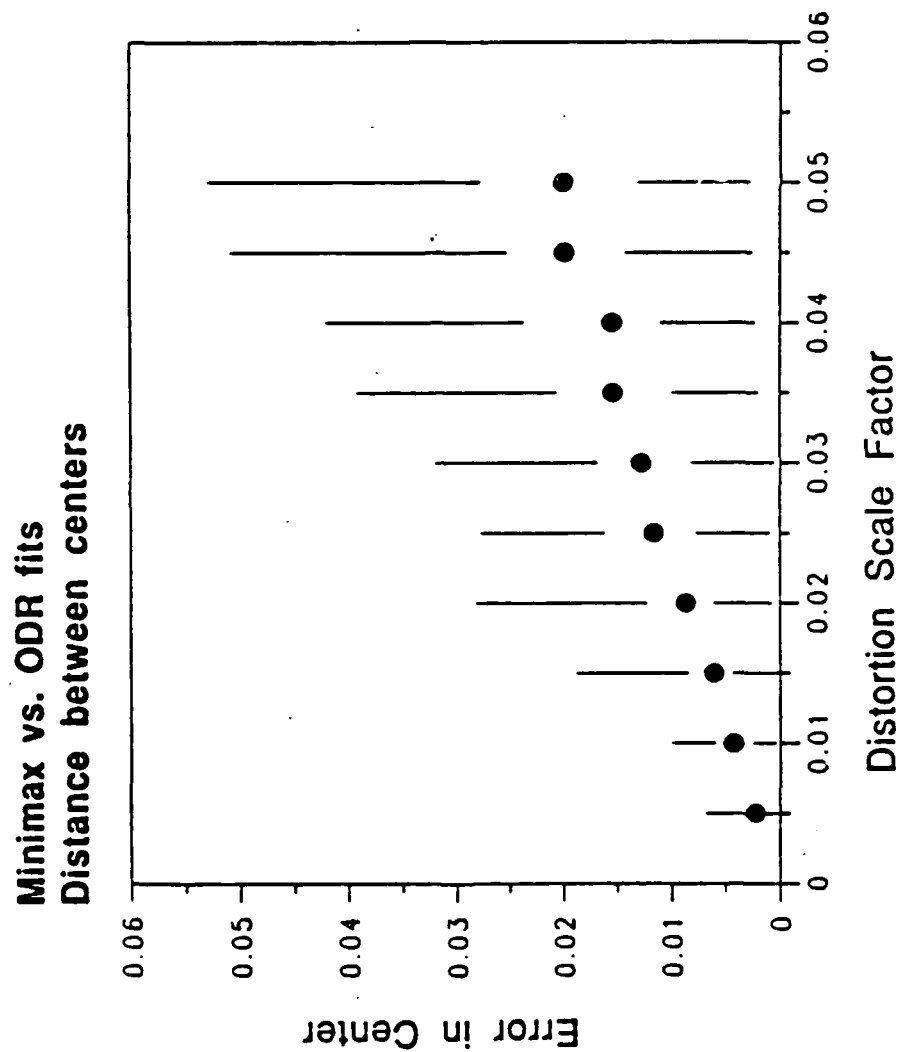


Errors in Sample Measurement Systems



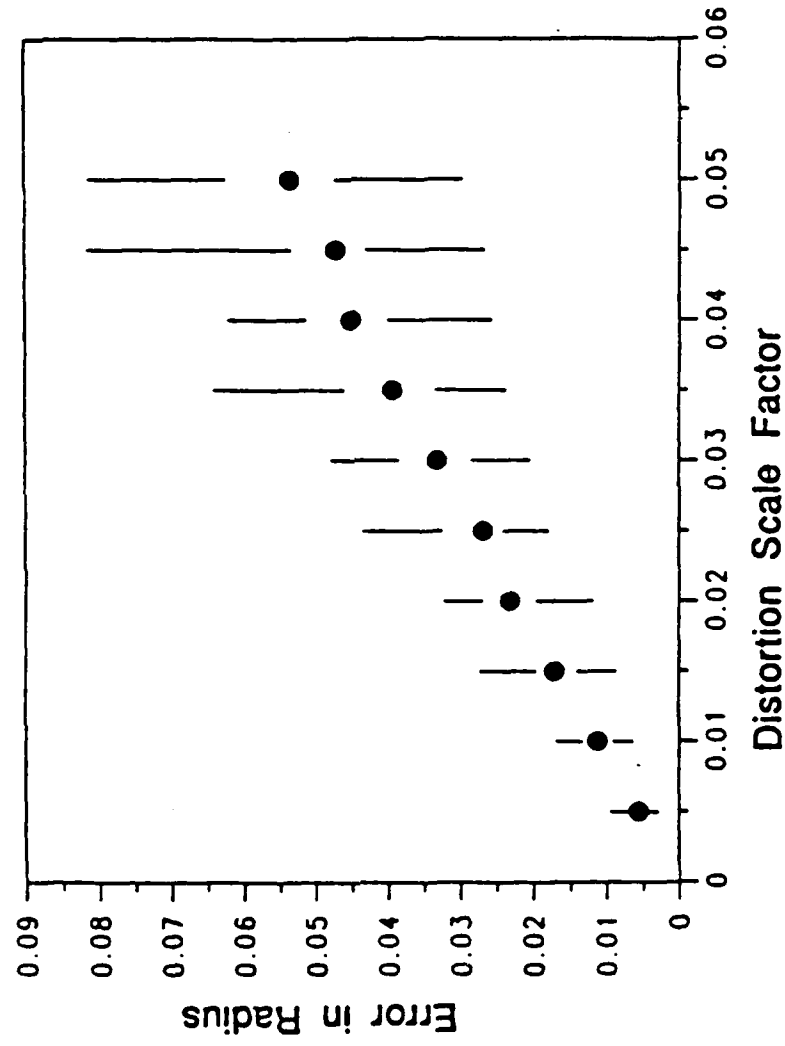
Experiment: Add errors to individual points. Investigate differences between fitting algorithms.

Errors in Sample Measurement Systems



Errors in Sample Measurement Systems

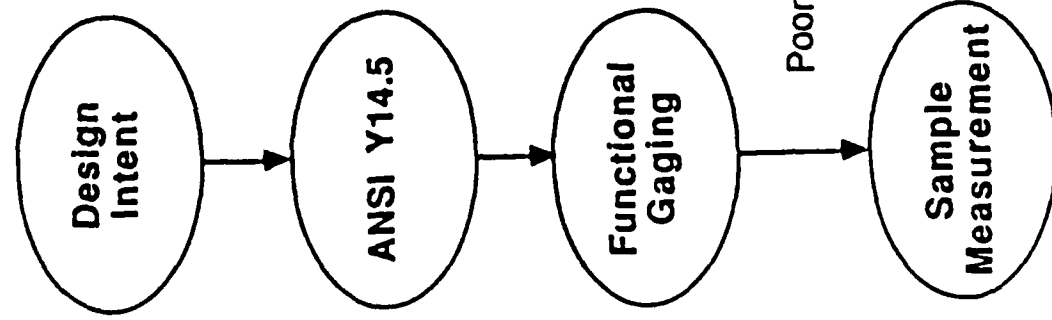
Minimax vs. ODR fits
Difference in radii



Quality Assurance Folklore, Scene 3

The Common View

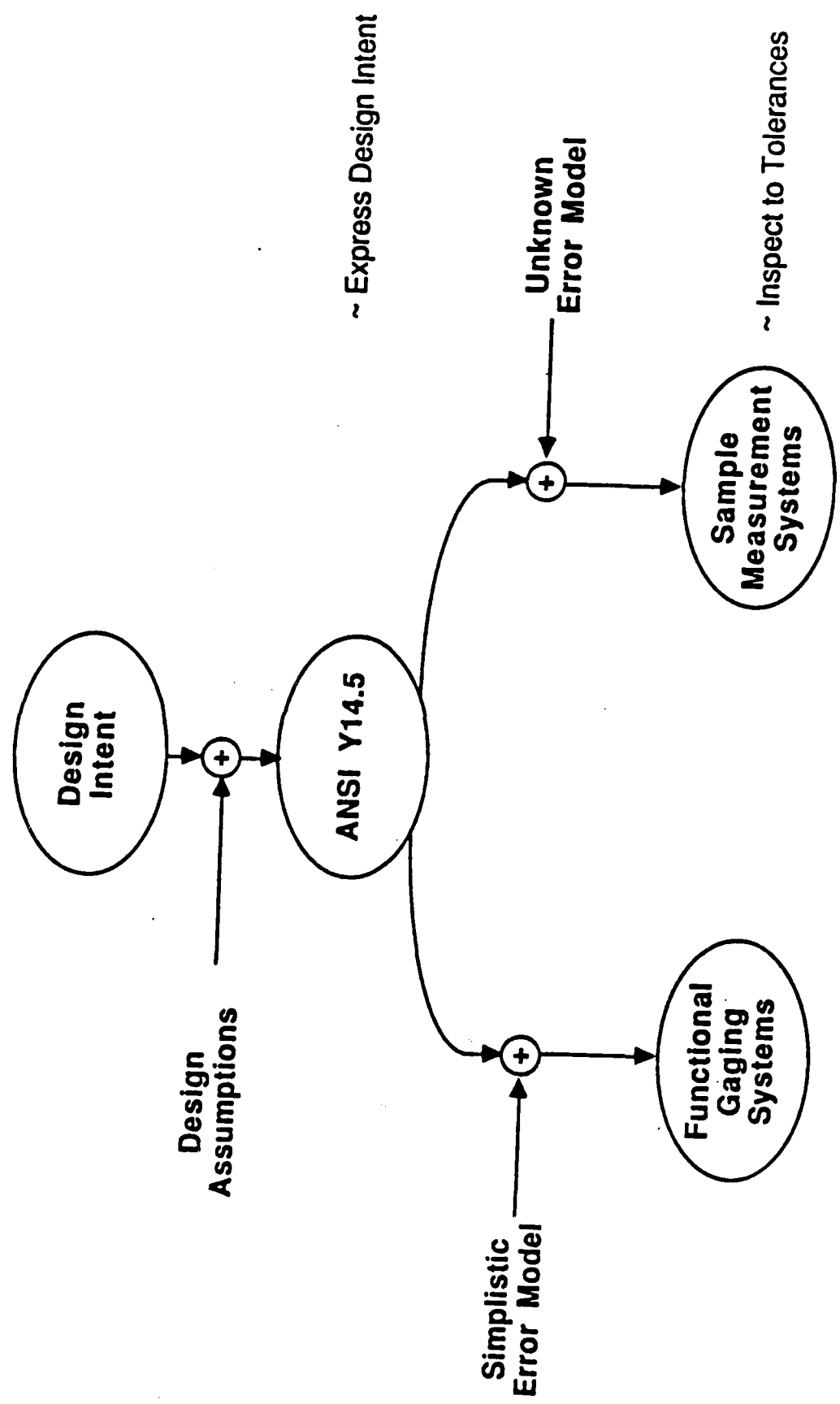
The Fallacy



Poorly limited by

The relationship
is irrelevant

Quality Assurance Reality, Scene 3



Components of the Problem

Cost of functional gages

No operational definition of tolerances for sample measurement systems

Small batches make statistical process control unstable

Inferiority complex of sample measurement community leads to adopting counterproductive model

Superiority complex of functional gaging community leads to overconfidence

Unrealistic models of the effects of gage tolerances

Sources of error in sample measurement systems are uncharacterized

The Solution

- Adopt a better model of quality assurance practice
- Develop more realistic models of functional gaging system errors
- Develop models of sample measurement system errors
- Develop improved usage guidelines and standards for each class of system

Together these steps will provide a rational basis for economic decision-making

A BAYESIAN PERSPECTIVE ON "TOLERANCING"

N.D. SINGPURWALLA

PREAMBLE:

MY REVIEW OF THE LITERATURE ON
"STATISTICAL TOLERANCING" LEADS ME TO
BELEIVE THAT THERE MAY BE A LACK
OF UNIFORMITY WITH REFERENCE TO
THE MEANING AND INTERPRETATION
OF THE NOTION OF TOLERANCE.

THE ABOVE CONSIDERATION LEADS
ME TO PROPOSE THE FOLLOWING AS A
STREAMLINED APPROACH TO VIEWING
TOLERANCE. IS MY APPROACH WELL KNOWN.

WE ATTEMPT TO MODEL THE DESIGN ENGINEER'S PERSPECTIVE.

- LET D BE SOME DIMENSION OF INTEREST TO A DESIGNER \mathcal{D} .
EX. D = DIAMETER OF A BOLT.
- SINCE THE NOTION OF A MEASUREMENT, SUCH AS A LENGTH, WIDTH, CIRCUMFERENCE, ETC., IS AN ABSTRACTION (IT EXISTS ONLY IN OUR MINDS AND SO CAN NEVER BE REALIZED IN REAL LIFE), D IS AN UNKNOWN QUANTITY - RESIDING IN \mathcal{D} 'S MIND.
- ALL UNCERTAINTY IS BEST DESCRIBED BY PROBABILITY, AND SO

MAY SUBJECTIVELY SPECIFY HIS/HER
UNCERTAINTY ABOUT REALIZING D VIA A PROB
STATEMENT (ASSUMED SYMMETRIC)

$$(1) \quad P\{d - \epsilon \leq D \leq d + \epsilon\} = 1 - \alpha,$$

WHERE $\alpha > 0$ IS A SMALL NO.

(1) ABOVE IS CALLED A "DESIGN SPEC."

[IT'S FREQUENTIST INTERPRETATION IS
THAT $(1 - \alpha)\%$ OF ITEMS PRODUCED WILL
HAVE $D \in [d + \epsilon, d - \epsilon]$].

d IS CALLED THE NOMINAL VALUE, AND

ϵ IS CALLED A TOLERANCE.

$d + (-)\epsilon$ IS CALLED AN UPPER (LOWER)
SPECIFICATION LIMIT.

• OBJECTIVE IS TO SPECIFY ϵ GIVEN α AND d .

THE MANUFACTURING SCENARIO

- LET D' BE THE DIMENSION CORRESPONDING TO D OF THE MANUFACTURED ITEM.
- ONCE AGAIN, D' IS AN UNKNOWN QUANTITY TO Θ , WHO DESCRIBES HIS/HER UNCERTAINTY ABOUT D' , VIA SAY A GAUSSIAN DISTRIBUTION OF THE FORM

$$(2) (D' | \mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2).$$

- SP. FURTHER THAT Θ IS UNCERTAIN ABOUT μ AND FOR Θ , $(\mu | m, \tau^2) \sim \mathcal{N}(m, \tau^2)$.

• IT NOW FOLLOWS (FROM STANDARD ARGUMENTS
THAT, FOR Θ ,

$$(3) \quad (D' | m, \sigma^2, \tau^2) \sim \mathcal{N}(m, (\sigma^2 + \tau^2)).$$

NOTE: IN THE LITERATURE, DISTRIBUTIONS
SUCH AS (3) ARE GIVEN A FREQUENTIST
INTERPRETATION, AND ARE VIEWED AS BEING
GENERATED AS THE ACTUAL CAPABILITIES OF
THE MANUFACTURING PROCESS. WE TAKE EXCEPTION
WITH THIS POINT OF VIEW AND REGARD (3) AS
 Θ 'S SUBJECTIVE ASSESSMENT ABOUT THE
UNCERTAINTY ABOUT D' . THE FREQUENTIST
INTERPRETATION OF (3) FAILS TO HOLD WHEN
WE CONSIDER FLEXIBLE MANUFACTURING SCENARIOS.

FROM (3) ABOVE, IT FOLLOWS, THAT FOR
 θ .

$$(4) \Pr \{ m - k(\sigma^2 + \tau^2)^{1/2} \leq D' \leq m + k(\sigma^2 + \tau^2)^{1/2} \} \\ = 1 - \beta, \text{ WHERE FOR ANY SPECIFIED } \beta, \\ k \text{ IS KNOWN.}$$

- $m \pm k(\sigma^2 + \tau^2)^{1/2}$ ARE KNOWN AS THE
"NATURAL TOLERANCE LIMITS" - AN
UNFORTUNATE CHOICE OF TERMINOLOGY - AND
- $m + k(\quad)^{1/2}$ [$m - k(\quad)^{1/2}$] IS KNOWN
AS THE UPPER (LOWER) TOLERANCE LIMIT.

THUS TO SUMMARIZE WE HAVE, AS A MODEL FOR Θ 'S THOUGHT PROCESSES:

$$(1) P_{\pi} \{ d - \epsilon \leq D \leq d + \epsilon \} = 1 - \alpha$$

FOR THE DESIGN PART, AND

$$(2) P_{\pi} \{ m - k(\sigma^2 + \tau^2)^{1/2} \leq D' \leq m + k(\sigma^2 + \tau^2)^{1/2} \} = 1 - \beta, \text{ FOR THE MANUFACTURING PART.}$$

WE PROPOSE THE FOLLOWING AS (PLAUSIBLE) THINK PIECES:

TOLERANCE

- A DESIGN IS A PRIORI GOOD IF $d = m$ AND $k(\sigma^2 + \tau^2)^{1/2} \leq \epsilon$ WHEN $\alpha \leq \beta$.
- A DESIGN IS A PRIORI TOLERANCE OPTIMAL IF $d = m$ AND $k(\sigma^2 + \tau^2)^{1/2} = \epsilon$ WHEN $\alpha = \beta$.

UPDATING \mathcal{D} 'S BELIEF ABOUT D' — POSTERIOR ANALYSIS.

- SUPPOSE THAT \mathcal{D} HAS HAD AN OPPORTUNITY TO OBSERVE THE MANUFACTURING PROCESS, AND AS A RESULT OBTAIN AS DATA

$\underline{d} = (d'_1, d'_2, \dots, d'_n)$ WHERE d'_i IS A REALIZATION OF D'_i , $i=1, \dots, n$.

- LET $\bar{d} = \sum d'_i / n$.

THEN, THE PREDICTIVE DISTRIBUTION OF D' GIVEN \underline{d} IS

$$p(D' | \underline{d}) = \int_{\text{all } \mu} p(D' | \mu) p(\mu | m, \tau^2, \underline{d}) d\mu$$
$$= N\left(\left(\frac{n\bar{d}\tau^2 + m\sigma^2}{n\tau^2 + \sigma^2}\right), \left(\sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\right)\right)$$

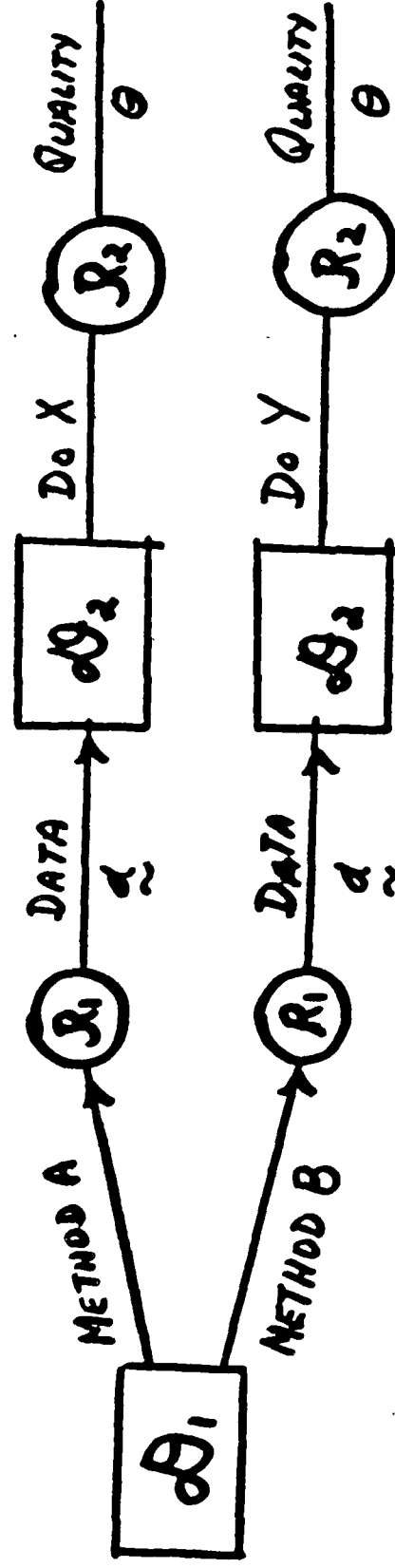
THUS, A DESIGN IS A POSTERIORI
TOLERANCE GOOD IF :

$$d = \frac{n\bar{d}T^2 + m\sigma^2}{nT^2 + \sigma^2} \quad \text{AND}$$

$$k \left(\sigma^2 + \frac{\sigma^2 T^2}{nT^2 + \sigma^2} \right) \leq \epsilon \quad \text{WHEN } d \leq \beta,$$

AND IT IS A POSTERIORI TOLERANCE
OPTIMAL IF THE INEQUALITIES ABOVE
ARE REPLACED BY EQUALITIES.

OUTLINE OF A RATIONAL BASIS FOR CHOOSING AMONG ALTERNATIVE TECHNOLOGIES FOR Q.A.



UTILITY $(\theta, x, \tilde{d}, A)$

EXPECTED UTILITY $U(x, \tilde{d}, A) = \int U(\theta, x, \tilde{d}, A) p(\theta) d\theta$

EXPECTED UTILITY $U(A) = \int U(x, \tilde{d}, A) p(\tilde{d}) d\tilde{d}$

CHOOSE THAT UTILITY WHICH MAXIMIZES EXPECTED UTILITY

$U(A)$

Session IV.

Uncertainties in
Design Optimization

PROBABILISTIC OPTIMAL DESIGN USING
SECOND MOMENT CRITERIA

by

Ashok D. Belegundu
Mechanical Engineering
The Pennsylvania State University
University Park, PA 16802

2

MOTIVATION

1. Emphasis on Deterministic Optimization

2. Robust Design

-- Variations in Loading, Dimensions, Material, Shape

* important while optimizing

-- Nonlinear Functions

* finite element modeling

* power transmissions; dynamic systems

* machine tolerances

3. LARGE PROBLEMS

$N_{DV} \geq 1500$, $N_{CON} \geq 5000$

PREVIOUS WORK

STRUCTURAL PROBLEMS

1967 F. MOSES, D. E. KINSEY

-- CONSIDERED CORRELATION

-- LOADS, MATERIAL

1971 VANMARCKE

1977 DAVIDSON, FELTON, HART (INCL. DYNAMIC LOADS)

1984 BRANDT, JENDO, MARKS

1985 FRANGOPOL

1986 THOFT-CHRISTENSEN, MUROTSU

1986 JEBENS (In German)

1986 JOZWIAK

-- TEXTS BY SIDDALL, HANGEN, RAO, S. S., SANDLE

1986 at U. Missouri / Taguchi Concepts

1987 BELEGUNDU

KINEMATICS (TOLERANCES)

1969 GARRET, HALL

1973 DHANDE, CHAKRABORTY

1978 A. C. RAO

1983 SHARFI, SMITH

1986 S. S. RAO

1986 T. S. MRUTHYUNJAYA

TOLERANCING: GROSSMAN ('76), TURNER ('87), LEE and T. C. WOR
(1986-87), TAGUCHI ('85)

SYSTEM DESIGN / QUALITY CONTROL: TAGUCHI
-- LOW SENSITIVITY

PREVIOUS WORK -- TREATMENT OF NONLINEAR FUNCTIONS

MEAN POINT APPROXIMATION

$$g(\underline{x}) \approx g(\underline{v}) + \left. \frac{dg}{dx_i} \right|_{\underline{v}} \cdot (x_i - v_i)$$

$$E(g) \approx g(\underline{v})$$

$$\sigma_g \approx \sqrt{\sum_i \left. \frac{dg}{dx_i} \right|_{\underline{v}}^2 \sigma_{x_i}^2}$$

Design Requirement

$$Pr(g \leq 0) \geq P_{s0}$$



$$\frac{-g(\underline{v})}{\sqrt{\sum_i \left. \frac{dg}{dx_i} \right|_{\underline{v}}^2 \sigma_{x_i}^2}} \geq \beta_0, \beta_0 = \Phi^{-1}(P_{s0})$$

* GIVES DIFFERENT ANSWERS FOR

$$g - \bar{g} \leq 0 \text{ and } \frac{g}{\bar{g}} - 1 \leq 0$$

Note:

* SOMETIMES GIVES SIGNIFICANT ERROR IN P_s .
- BUILT EFFECT ON DESIGN?

5

PROBABILISTIC ANALYSIS USING SECOND-MOMENT CRITERIA

. Hasofer-Lind (H-L) Reliability Index

. Basis for generalised reliability indices

CONSIDER

$$G = G(\underline{x})$$

\underline{x} = random variable vector, $(n \times 1)$,

DEFINE

$$d_i = \frac{x_i - v_i}{\sigma_i}$$

v_i = mean of x_i

σ_i = std. dev. of x_i

THUS

$$G(x_i) \longrightarrow g(\sigma_i d_i + v_i)$$

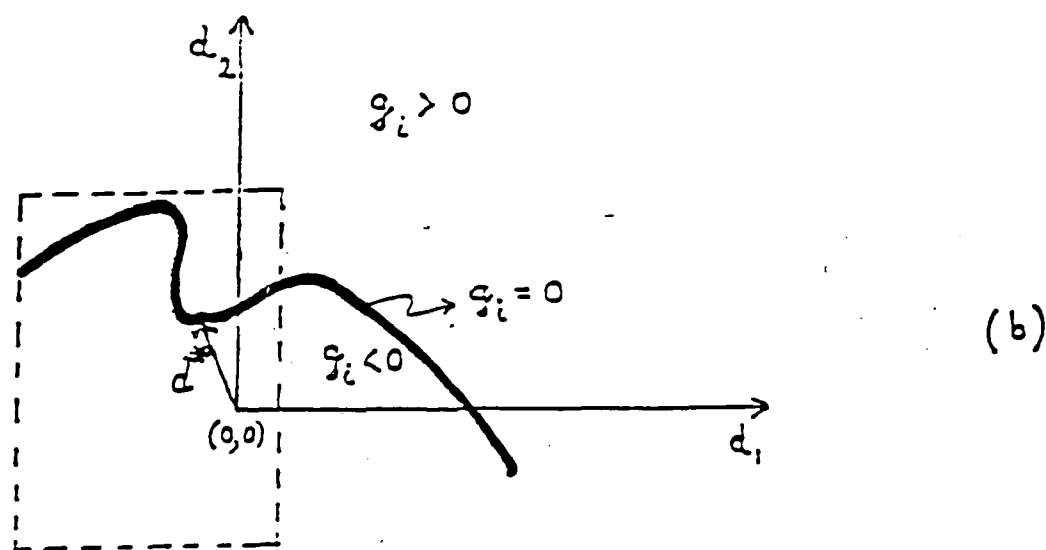
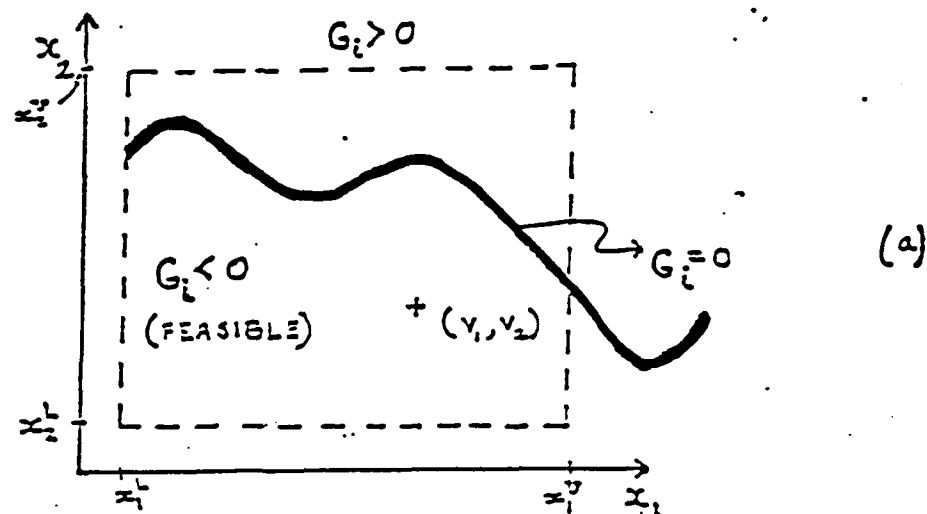
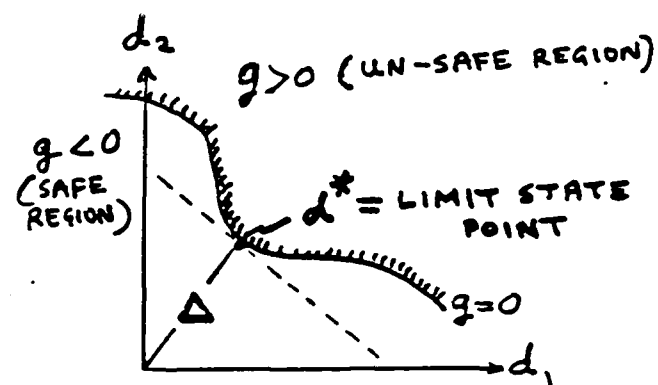


Figure 1. (a) Constraint in Original Space, (b) Constraint in Reduced Space.

LIMIT STATE

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \underline{d}^T \underline{d} \\ & \text{subject to} \quad g = 0 \end{aligned}$$



$$X_i^* = \sigma_i d_i^* + \mu_i$$

$$\text{SHORTEST DISTANCE: } \Delta = \sqrt{\underline{d}^{*T} \underline{d}^*} \cdot (-\text{sgn}(g)_{d=0})$$

Design Requirement:

$$\Pr(G \leq 0) \geq P_{s0}$$

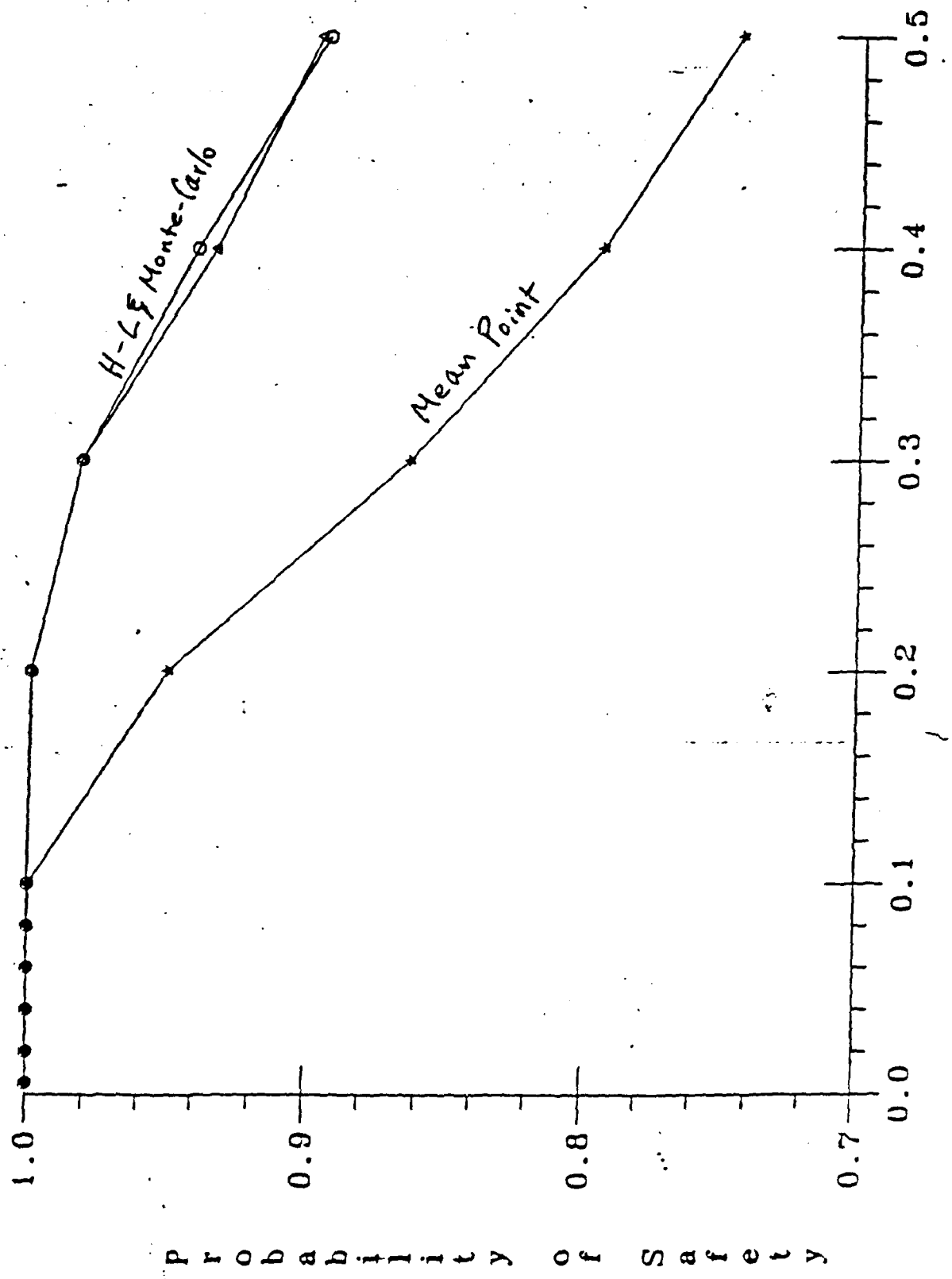
becomes[†]

$$\Delta \geq \beta_0$$

[†] exact if G is linear and x is normal

Spring Frequency Constraint $g(d,D) = 1 - \frac{14845d}{D^3 n}$ $d,D > 0$ $n = 1991$

$u_d = 0.071$ $u_D = 1.5$



Coefficient of Variance

Algorithm for the sub-problem to find Δ

$$\begin{array}{ll} \text{minimize} & 1/2 \underline{d}^T \underline{d} \\ \text{subject to} & g(\underline{d}) = 0 \end{array}$$

Step 1 Choose initial design \underline{d}^0

Set $k=0$

Step 2 Solve QP

$$\text{Minimize } 1/2 \underline{p}^T \underline{p} + \underline{d}^{k^T} \underline{p}$$

$$\text{Subject to } \nabla g^k \underline{p} + \underline{g}^k = 0$$

Let \underline{p}^k be the solution.

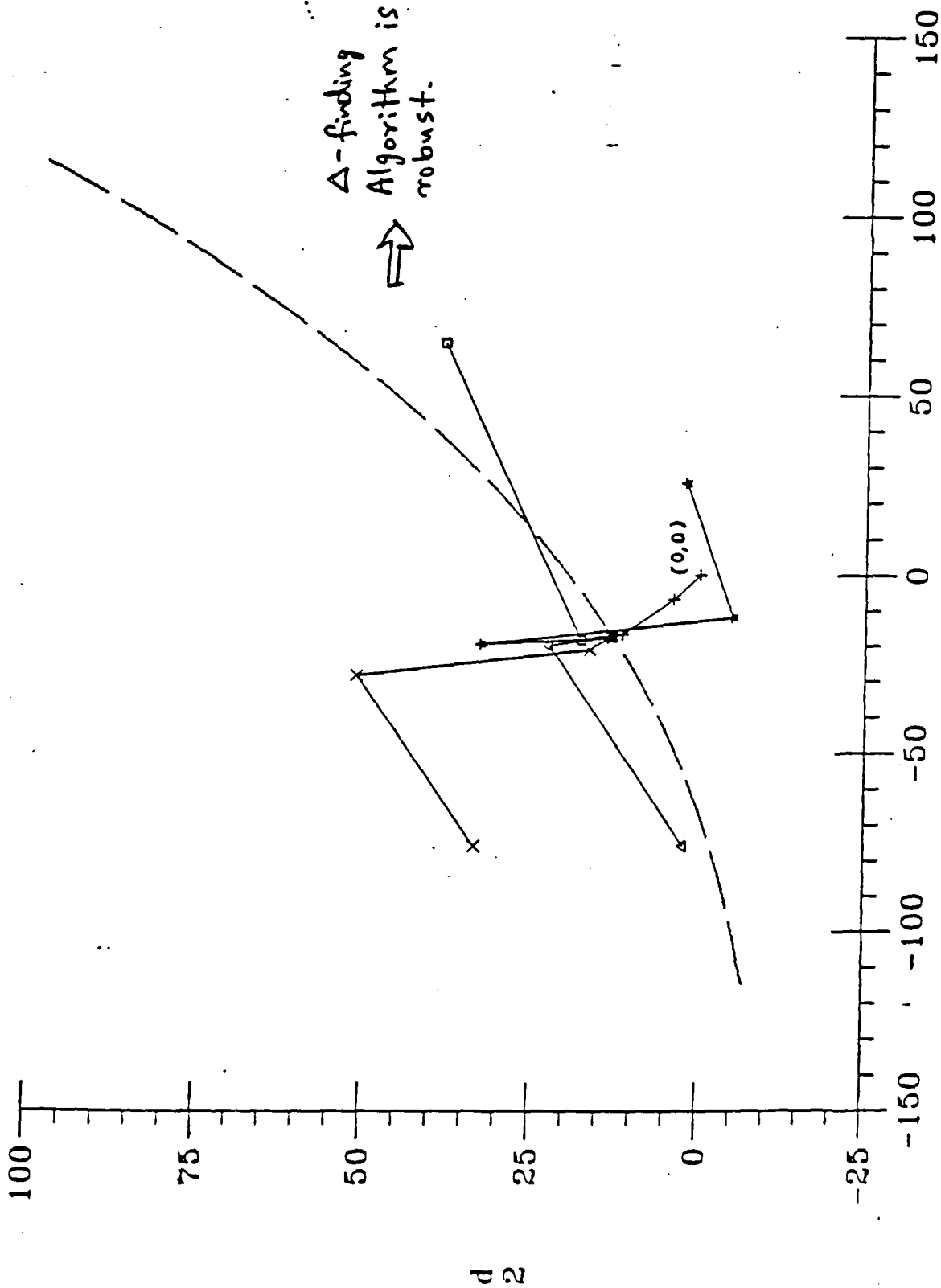
Step 3 Set $\underline{d}^{k+1} = \underline{d}^k + \alpha \underline{p}^k$

Step 4 If $|g(\underline{d}^k)| \leq TOL$, and $\|\underline{p}^k\| \leq TOL$,

Then set $\Delta = (\sum_{i=1}^n d_i^2)^{1/2}$, and stop.

Otherwise, set $k=k+1$ and goto step 2.

Dash Line : $g(d_1, d_2) = 0$ in Reduced Space



d_1

IMPLICATIONS OF LOW SENSITIVITY

THE H-L CRITERION

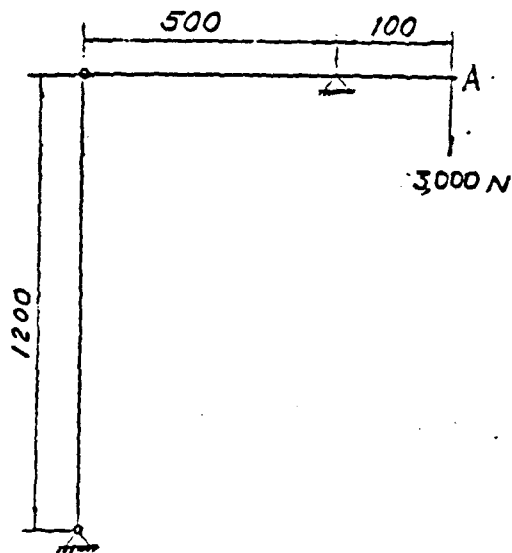
$$\Delta \geq \beta_0$$

CAN BE WRITTEN AS

$$\frac{- \sum_i \left. \frac{dg}{dx_i} \right|_{x^*} \cdot (x_i^* - v_i)}{\sqrt{\sum_i \left. \frac{dg}{dx_i} \right|_{x^*}^2 \sigma_i^2}} \geq \beta_0$$

NOTE: $\left. \frac{dg}{dx} \right|_{x^*} \uparrow \rightarrow \Delta \downarrow$

EXAMPLE ILLUSTRATING LOW-SENSITIV



$$E_1 = 200,000 \text{ N/mm}^2$$

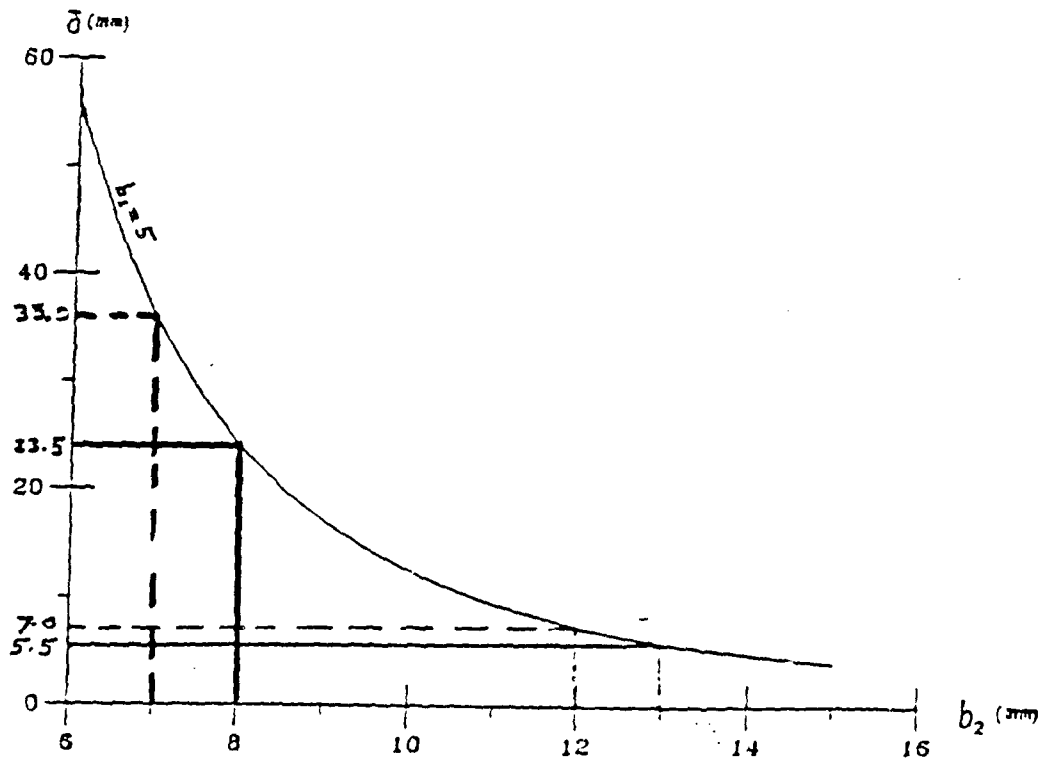
Unit : mm

Displacement at end A

$$\delta = \frac{0.2057}{b_1} + \frac{12000}{b_2^3}$$



$$E_2 = 70,000 \text{ N/mm}^2$$



A PROBABILISTIC OPTIMAL DESIGN PROBLEM WITH H-L INDEX

$$\text{Minimize}_{\underline{v}^d} \quad E[f(x)] \approx f(\underline{v})$$

$$\text{Subject to} \quad \Delta_i \geq \beta_{0i} \quad i=1, \dots, m$$

$$\text{and} \quad \underline{v}^L \leq \underline{v}^d \leq \underline{v}^u$$

* ABOVE DOES NOT CONSIDER

(i) CORRELATION BETWEEN FAILURE MODES

(ii) COST DEPENDING ON P_s

* FORMULATION DEPENDS ON RISK ORIENTED ATTITUDE OF DESIGNER.

DESIGN SENSITIVITY ANALYSIS OF THE H-L RELIABILITY INDEX

Problem :
$$\left[\begin{array}{l} \text{minimize } \frac{1}{2} \underline{d}^T \underline{d} \\ \text{subject to } g = 0 \end{array} \right] \quad (1)$$

$$\underline{d}^* = \underline{d}^*(\underline{v})$$

$$\Delta = \Delta(\underline{v})$$

$$\boxed{\frac{d\Delta}{d\underline{v}^d} = ?}$$

$$\underline{x} = (x_1, x_2, \dots, x_k, \dots, x_n)^T$$

$$\underline{v} = (\underbrace{v_1, v_2, \dots, v_k}_{\underline{v}^d}, \dots, v_n)^T, \quad \underline{v}^d = \text{DESIGN VECTOR}$$

Sensitivity :

Necessary Conditions For (1) : $\underline{d}^* + \underline{N}^* \lambda^* = 0 \quad (2)$

$$g(\underline{d}^*, \underline{v}) = 0 \quad (3)$$

where

$$\underline{N}^* \triangleq \frac{dg^T(\underline{d}^*, \underline{v})}{d\underline{d}}$$

Differentiate (2) ~~and (3)~~ w.r.t. \underline{v}^d

$$\left[d\underline{d}^* / d\underline{v}^d \right] = - \left[\lambda^* d\underline{N}^* / d\underline{v}^d + \underline{N}^* d\lambda^* / d\underline{v}^d \right] \quad (4)$$

↑
OBTAIN
FROM (7)

SENSITIVITY cont'd.---

SUBSTITUTE FOR \underline{d}^* FROM (2) INTO (3):

$$g(-\lambda^* \underline{N}^*, \underline{v}) = 0 \quad (5)$$

DIFFERENTIATE (5) W.R.T. \underline{v}^d :

$$-\underline{N}^{*T} \left[\lambda^* \frac{d\underline{N}^*}{d\underline{v}^d} + \underline{N}^* \frac{d\lambda^*}{d\underline{v}^d} \right] + \frac{dg}{d\underline{v}^d} = 0 \quad (6)$$

↓

SOLVE
FIRST →

$$\frac{d\lambda^*}{d\underline{v}^d} = \left[\frac{dg}{d\underline{v}^d} - \lambda^* \underline{N}^{*T} \frac{d\underline{N}^*}{d\underline{v}^d} \right] / (\underline{N}^{*T} \underline{N}^*) \quad (7)$$

NOW SUBSTITUTE FOR $d\lambda^*/d\underline{v}^d$ FROM ABOVE
INTO (4) TO OBTAIN:

$$\checkmark \quad \frac{d\underline{d}^*}{d\underline{v}^d} = \text{SENSITIVITY OF } \underline{d}^* \text{ W.R.T. DESIGN}$$

and

$$\checkmark \quad \frac{d\Delta}{d\underline{v}^d} = \text{SENSITIVITY OF } \Delta \text{ W.R.T. DESIGN}$$

— x —

USE OF SECOND-ORDER DERIVATIVES

REQUIRE: \underline{N}^* and $d\underline{N}^*/d\underline{y}^d$.



$$\frac{dG}{dx_i}$$

and

$$\frac{d^2G}{dx_i dx_j}$$

FIRST-
ORDER

SECOND-
ORDER

FIRST-ORDER : USED COMMONLY IN OPTIMIZATION

SECOND-ORDER : NEEDED TO 'PUSH' THE DESIGN
INTO LOW-SENSITIVE REGIONS

- IMPORTANCE ESTABLISHED HERE
FOR PROBABILISTIC DESIGN
- EXPRESSIONS AVAILABLE
 - USED IN PAST TO
ACCELERATE CONVERGENCE,
THOUGH NOT ENTIRELY
SUCCESSFUL

FIRST- AND SECOND- ORDER DERIVATIVES IN STRUCTURAL RESPONSE

Finite Element Model

$$q = q(\underline{x}, \underline{z}) : \underline{K}(\underline{x}) \underline{z} = \underline{F}(\underline{x})$$

Direct Method

-- First-Order

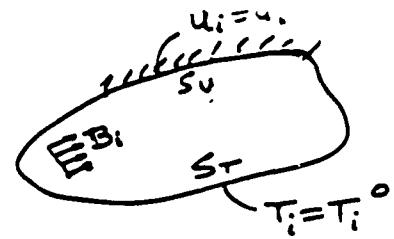
$$K \left[\frac{d\underline{z}}{d\underline{x}} \right] = - \frac{\partial}{\partial \underline{x}} [K(\underline{x}) \underline{z} - F(\underline{x})]$$

$$\frac{dq}{d\underline{x}} = \frac{\partial q}{\partial \underline{x}} + \frac{\partial q}{\partial \underline{z}} \cdot \left[\frac{d\underline{z}}{d\underline{x}} \right]$$

-- Second-Order

$$\begin{aligned} \frac{d^2 q}{d\underline{x} d\underline{x}_i} &= \frac{\partial^2 q}{\partial \underline{x} \partial \underline{x}_i} + \frac{d\underline{z}^T}{d\underline{x}_i} \frac{\partial^2 q}{\partial \underline{x} \partial \underline{z}} + \frac{\partial q}{\partial \underline{z}} \cdot \frac{\partial}{\partial \underline{x}} \left(\frac{d\underline{z}}{d\underline{x}_i} \right) \\ &+ \frac{\partial^2 q}{\partial \underline{z} \partial \underline{x}_i} \cdot \frac{d\underline{z}}{d\underline{x}} + \frac{d\underline{z}^T}{d\underline{x}_i} \cdot \frac{\partial^2 q}{\partial \underline{z}^2} \cdot \frac{d\underline{z}}{d\underline{x}} \\ &+ \frac{\partial q}{\partial \underline{z}} \cdot \frac{\partial}{\partial \underline{z}} \left(\frac{d\underline{z}}{d\underline{x}_i} \right) \frac{d\underline{z}}{d\underline{x}} \end{aligned}$$

Adjoint Method : ✓



Continuum Model

$$G = \int_V [\psi(\phi_k, \sigma_{ij}) + h(\phi_k, u_i)] dV + \int_{S_u} F(T_i) dS + \int_{S_T} g(u_i) dS$$

$$\sigma_{ij,j} + B_i = 0 \quad \text{in } V$$

$$\sigma_{ij} n_j = T_i \quad \text{in } S$$

$$\sigma_{ij} = [D] \epsilon_{kl}$$

$$\epsilon_{ij} = (u_{i,j} + u_{j,i})/2$$

WHERE:

$$\phi_k = \phi_k(x, y, z) = \text{DESIGN FUNCTIONS}$$

SENSITIVITY: ✓

- Shape Variations -- stochastic boundary
- Equivalent to First-Order and Second-Order "Perturbation Analysis"
- Dynamic Systems ✓

variation are taken as $\beta_1 = \beta_2 = \beta_3 = 0.2$. The reliability index $\beta_1 = \beta$ is taken as same for all constraints. The material weight density is 0.1 lb/in³, Young's modulus $E = 10^7$ psi, allowable stresses are $\sigma_1^a = \sigma_2^a = 5000$ psi, $\sigma_3^a = 20,000$ psi, displacement limits are $\delta_1^a = \delta_2^a = 0.005$ in., lower bound in natural frequency equals 2178 Hz, and a lower bound of 0.05 in² is imposed on each design variable. The starting design is $v = (10.0, 5.0, 5.0)$ in². Many other starting designs were used and resulted in the same solutions.

Table 2 contains results for various values of β .

TABLE 2
RESULTS FOR 1-BAR TRUSS

β	FINAL COST	FINAL DESIGN	ACTIVE SET	SENSITIVITY OF ACTIVE CONSTRAINTS		
0 (deterministic)	10.542	9.71 1.33 0.23	2, 3	—	—	—
0.05	10.740	9.70 2.21 0.21	2, 3	-10.4 -3.1 -17.1	0.1 -19.1	-9.4
0.1	10.919	9.75 2.18 0.20	2, 3	-5.1 -1.6 -12.4	1.0 -9.6	-1.6
0.2	11.211	9.33 1.83 0.23	2, 3	-2.3 -0.9 -6.8	0.9 -4.4	-1.1
0.3	11.455	9.90 2.10 0.20	1, 2	1.3 -1.6 0.6	-0.7 -0.8	-1.8
0.4	11.522	9.39 1.34 0.14	1, 2	1.0 -1.1 0.6	-0.3 -0.6	-1.9
0.5	11.193	10.10 1.48 0.22	1, 3	0.7 -0.9 0.3	-0.4 -0.3	-1.5
0.6	11.924	10.37 1.87 0.12	1, 3	0.6 -0.7 0.1	-0.1 -0.4	-1.1
0.7	14.272	10.47 1.79 0.73	1, 2	0.3 -0.6 0.17	-0.23 -0.3	-1.0
0.8	14.920	10.55 1.77 0.31	1, 2	0.4 -0.5 0.14	-0.1 -0.23	-0.8

NOTE: 1 = natural frequency
Active constraint no. 1 = natural frequency
2 = horizontal displacement, load case 1
3 = vertical

It may be noted that the deterministic solution is obtained by solving problem (36) and corresponds, in effect, to a value of $\beta = 0$. The design sensitivity analysis expressions derived in this paper have been verified using a divided-difference scheme based on

$$(38) \quad dg/db_i = [g(v_1, \dots, v_i + \epsilon, \dots, v_k) - g(v_1, \dots, v_k)]/\epsilon$$

where ϵ is a small number. The sensitivity of the active second-moment criteria, d/dv_i ($i = 1, \dots, 3$) are printed out in Table 2. The results in Table 2 provide the designer with a choice of designs, depending on the desired reliability level.

Firstly, note that no solution exists for $\beta \geq 1$. Since the coefficient of variation α is same for each variable, the results only depend on the product $\alpha\beta$. Thus, the absence of a solution for $\alpha = 0.2$ and $\beta \geq 1$ also holds true for $\alpha = 0.1$ and $\beta \geq 2$. Thus, for a specified value of α , the maximum reliability of the structure is known.

Secondly, note the sudden change in the active set when going from $\beta = 0.2$ to $\beta = 0.3$. Based on the

previous remark, we can also conclude that for a fixed $\beta = 1$, there is a sudden change in the governing failure modes when α is increased from 0.04 to 0.06. This also substantiates the fact that considering only the failure modes based on a deterministic design is generally unsafe; unexpected failure modes can play a role in probabilistic design.

Thirdly, it is interesting to conduct parametric studies related to β . Specifically, changes in optimum cost and changes in the sensitivity values with respect to β may provide the designer with additional insights.

Tension-Compression Spring (Fig. 3)

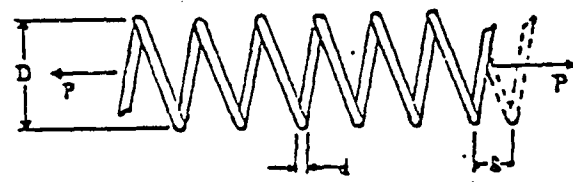


Figure 3. Tension-Compression Spring.

The problem [16] is to minimize the (expected) weight of the spring subject to constraints on minimum deflection, shear stress, surge frequency, limits on outside diameter and on design variables. The wire diameter d , coil diameter, D and number of coils n have been considered as random variables, and their mean values are considered as design variables. After accounting for the input data, the deterministic problem can be expressed as

Minimize $(n + 2) D d^2$

Subject to

$$1 - \frac{\beta^3 n}{7.1875 \times 10^{-3} d^3} \leq 0$$

$$\frac{\beta^2 - d^2}{1.2566 \times 10^{-3} (D d^3 - d^4)} + \frac{1}{5.108 \times 10^3 d^2} - 1 \leq 0$$

$$(39) \quad 1 - \frac{180.35 d}{\beta^2 n} \leq 0$$

$$\frac{D + d}{1.5} - 1 \leq 0$$

$$d \geq 0.05$$

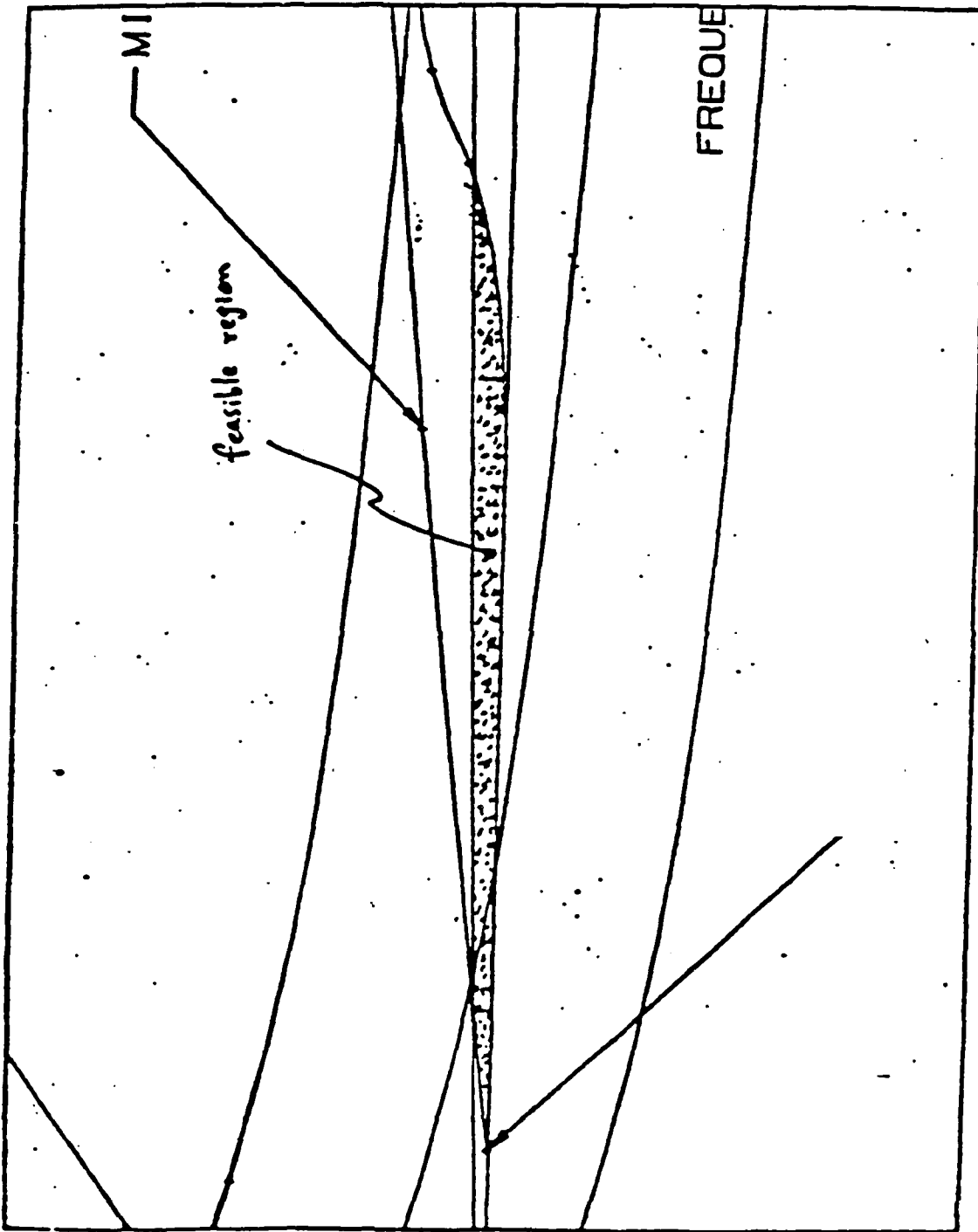
$$D \geq 0.05$$

$$n \geq 1$$

The POD problem corresponding to (39) is in the form (37) with $s = 4$, $k = 3$. The results of the POD problem, for various values of β , are given in Table 3. For this problem $\alpha_1 = 0.005$, $\alpha_2 = 0.05$, $\alpha_3 =$

deterministic

'WITH 200M'



↑ A

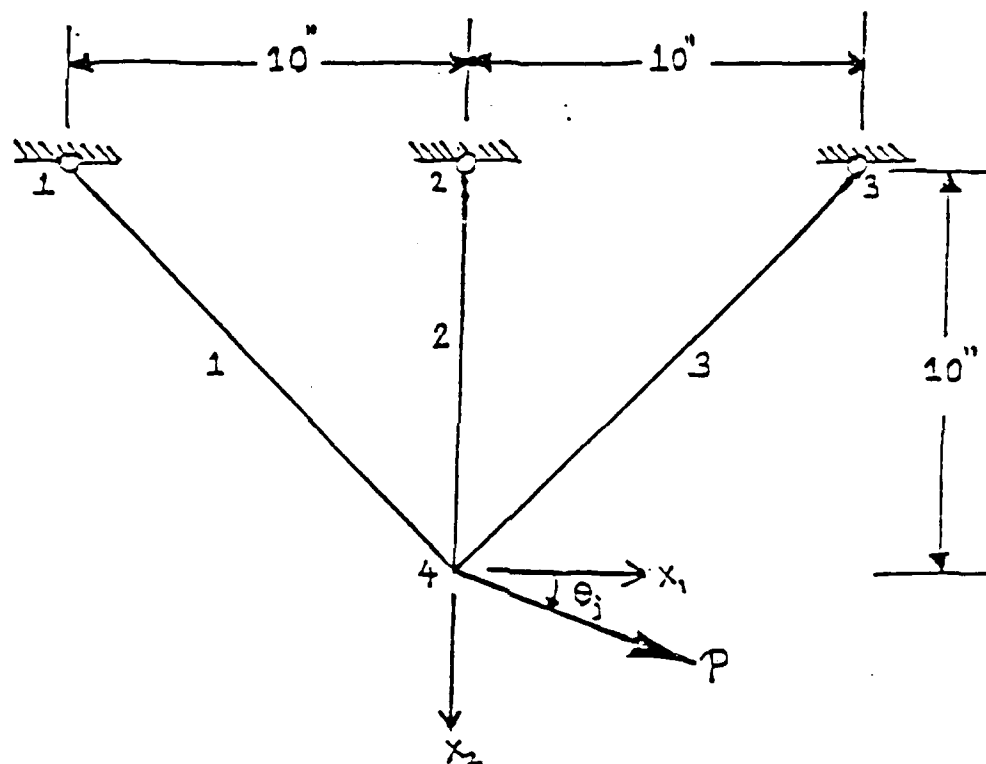


Figure 2. 3-Bar Truss.

SUMMARY & FUTURE WORK

- . General Approach for Probabilistic Optimal Design
 - robust design
- . Reliability Index Methods
- . Quality Control ; Tolerancing
 - "UP-FRONT"
- . Extensions : SYSTEM RELIABILITY
FPI
- . DYNAMIC SYSTEMS
- * . IDENTIFICATION OF PROBLEMS FOR WHICH
UNCERTAINTY OR VARIATION IS CRITICAL.

TABLE 1
LOAD CONDITIONS FOR 3-BAR TRUSS

Load Case	1	2	3
θ , deg	45	90	135
P, kips	40	30	20

TABLE 2
RESULTS FOR 3-BAR TRUSS

$\Omega = (0.2, 0.2, 0.2)$						
δ	FINAL COST	FINAL DESIGN	ACTIVE SET	SENSITIVITY OF ACTIVE CONSTRAINTS		
0 (deterministic)	20.542	8.91 1.93 4.25	2,8	—		
0.05	20.768	9.30 2.21 3.82	2,8	-10.4 -3.2	4.1 -27.1	-5.4 -19.3
0.1	20.919	9.35 2.18 3.90	2,8	-5.1 -1.6	2.0 -13.4	-2.6 -9.4
0.2	21.211	9.33 2.03 4.23	<u>2,8</u>	-2.5 -0.9	0.9 -6.6	-1.1 -4.4
0.3	21.855	9.80 2.38 3.98	<u>1,2</u>	1.3 -1.6	-0.7 0.6	-2.8 -0.9
0.4	22.522	9.98 2.54 4.14	1,2	1.0 -1.1	-0.5 0.4	-2.0 -0.6
0.5	23.153	10.16 2.68 4.32	1,2	0.7 -0.9	-0.4 0.3	-1.5 -0.5
0.6	23.924	10.37 2.87 4.52	1,2	0.6 -0.7	-0.3 0.2	-1.2 -0.4
0.7	24.272	10.47 2.79 4.73	1,2	0.5 -0.6	-0.25 0.17	-1.0 -0.3
0.8	24.938	10.65 2.93 4.91	1,2	0.4 -0.5	-0.2 0.15	-0.8 -0.25

NOTE: $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$

Active constraint no. 1 = natural frequency

" " " 2 = horizontal displacement, load case 1

" " " 8 = vertical " " 2

TABLE 3
RESULTS FOR TENSION-COMPRESSION SPRING

$$\Omega = (.005, .05,)$$

B	FINAL COST X 0.001	FINAL DECISION	SENSITIVITY OF ACTIVE CONSTRAINTS		
0 (deterministic)	154.0	(.064, .750, 2.945)	-----		
0.05	156.0	(.065, .752, 2.971)	6813.0 -19241.0	-439.0 507.0	-37.0 0.
0.10	158.0	(.065, .757, 2.971)	3389.0 -9608.0	-218.0 252.2	-18.5 0.
0.2	162.0	(.065, .768, 2.973)	1677.0 -4791.0	-107.0 125.0	-9.2 0.
0.4	171.0	(.066, .789, 2.975)	821.0 -2283.0	-51.4 61.3	-4.5 0.
0.6	180.0	(.067, .812, 2.977)	536.0 -1580.0	-33.0 40.0	-3.0 0.
0.8	189.0	(.068, .834, 2.980)	393.5 -1179.0	-23.9 29.5	-2.3 0.
1.0	200.0	(.068, .858, 2.982)	308.0 -938.0	-18.4 23.1	-1.8 0.
1.2	211.0	(.069, .883, 2.985)	251.3 -777.0	-14.7 18.9	-1.5 0.
1.4	222.0	(.070, .908, 2.988)	211.0 -662.6	-12.2 15.9	-1.23 0.
1.65	237.0	(.071, .942, 2.992)	174.0 -558.0	-9.8 13.1	-1.0 0.

**DESIGN OPTIMIZATION WITH
INNER AND OUTER NOISE**

By

**Eric Sandgren
School of Mechanical Engineering
Purdue University**

OBSERVATION:

- We are not solving the design optimization problem incorrectly.
- We are solving the wrong design optimization problem

PROBLEM WITH TRADITIONAL METHODS:

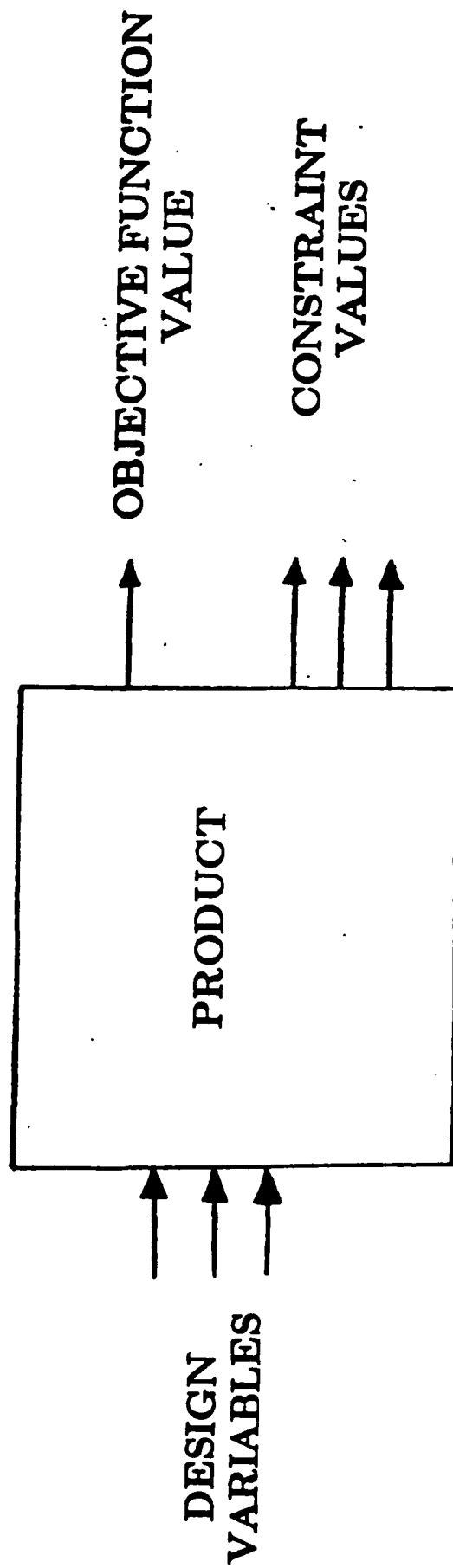
- Design is a fuzzy endeavor.
- Optimization requires a precise mathematical representation.

UNCERTAINTIES IN DESIGN:

- Specifications**
- Material properties**
- Loading conditions**
- Service environment**
- Manufacturing process**
- Idealized model**

INNER NOISE: Controllable variations in the design variables caused by time and manufacturing (wear and tolerances)

OUTER NOISE: Uncontrollable variations in design parameters (temperature, humidity)



STANDARD NONLINEAR PROGRAMMING
MODEL

GENERAL MATHEMATICAL PROGRAMMING PROBLEM

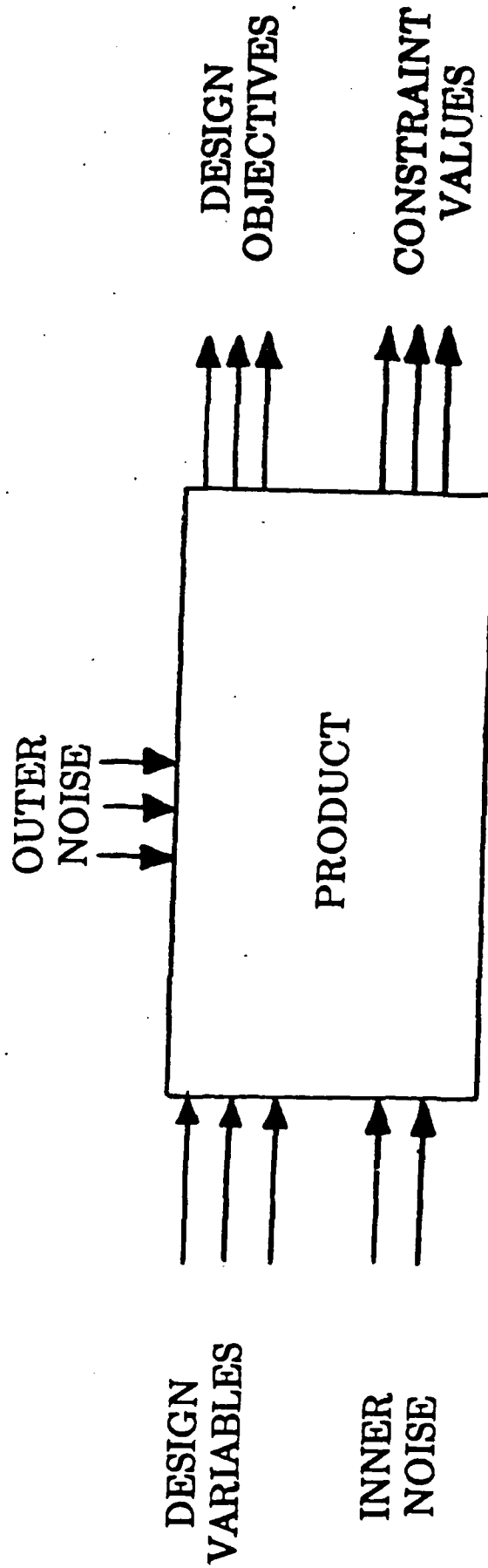
MINIMIZE $F(\bar{x}); \bar{x} = [x_1, x_2, \dots, x_N]^T, x \in \mathbb{R}^N$

SUBJECT TO

$$G_k(\bar{x}) \geq 0, k = 1, 2, 3, \dots, K$$

$$H_j(\bar{x}) = 0, j = 1, 2, 3, \dots, J$$

$$a_i \leq x_i \leq b_i \quad i = 1, 2, 3, \dots, N$$



THE REAL DESIGN OPTIMIZATION
PROBLEM

DESIGN QUESTIONS:

- Why is 50,000 PSI acceptable and not 50,001?
- Is $E = 30.0 \text{ E}+06$ or is it really $30.1+06$?
- Is $F_x = 10,000 \text{ lb}$ or is $F_x 10,100 \text{ lb}$?

METHODS FOR DEALING WITH UNCERTAINTY

- Monte Carlo Simulation**
- Stochastic Programming**
- Fuzzy Optimization**
- Design for Latitude**
- Game Theory (Multiple Objectives)**

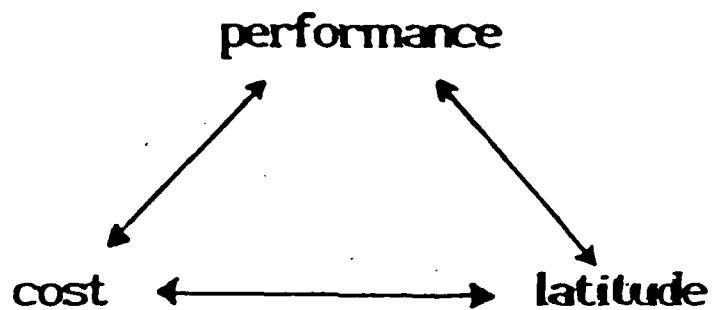


Figure 2: The Trade-off Triangle

EXAMPLE PROBLEM FOR GOAL PROGRAMMING FORMULATION

DIFFERENTIAL RATIO (x_1)

BORE (x_2)

STROKE (x_3)

- 1) THE EPACITY ECONOMY WILL BE GREATER THAN 30 MPG.
- 2) THE (0 - 60) MPH TIME SHOULD BE LESS THAN 11 SECONDS.
- 3) THE TOP SPEED SHOULD BE AS CLOSE AS POSSIBLE TO 95 MPH.

$$1) a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + d_1^- - d_1^+ = 30 - \text{Initial EPACITY}$$

$$2) a_{21} x_1 + a_{22} x_2 + a_{23} x_3 + d_2^- - d_2^+ = 11 - \text{Initial T(0-60)}$$

$$3) a_{31} x_1 + a_{32} x_2 + a_{33} x_3 + d_3^- - d_3^+ = 95 - \text{Initial VMAX}$$

$$x_1 \Rightarrow x_1^+ - x_1^-$$

$$a_{ij} = \frac{\text{GPSIM Result } (x_j + \Delta x_j) - \text{GPSIM Result } (x_j)}{\Delta x_j}$$

$$z = P_1 d_1^- + P_2 d_2^+ + P_3 (d_3^- + d_3^+)$$

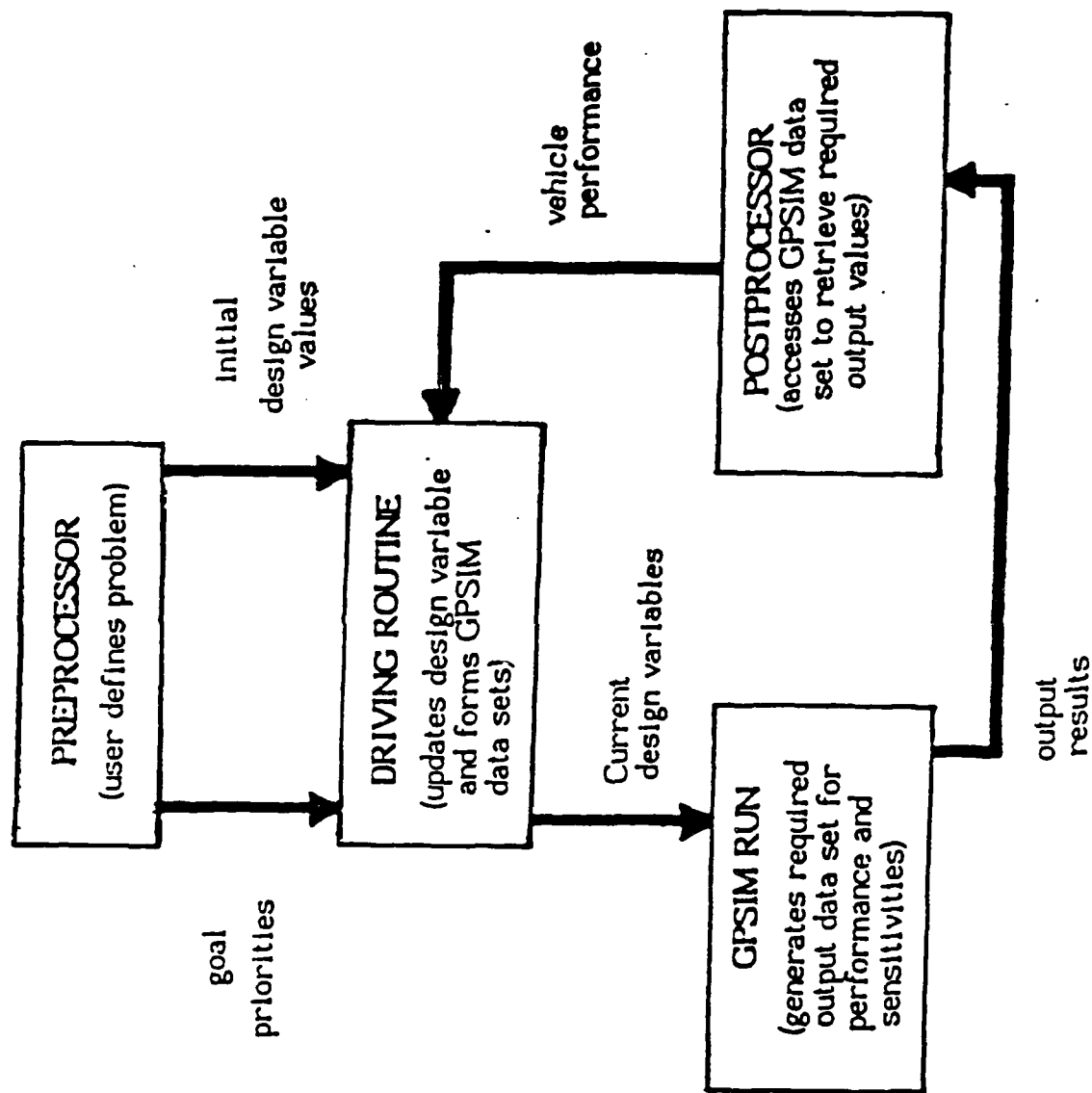


FIGURE 1: Program Flow for Conceptual Powertrain Optimization Algorithm

THE REAL DESIGN PROBLEM:
OBJECTIVES (GOALS)

- Minimize weight
- Maximize natural frequency
- Minimize chance of buckling
- Minimize maximum stress

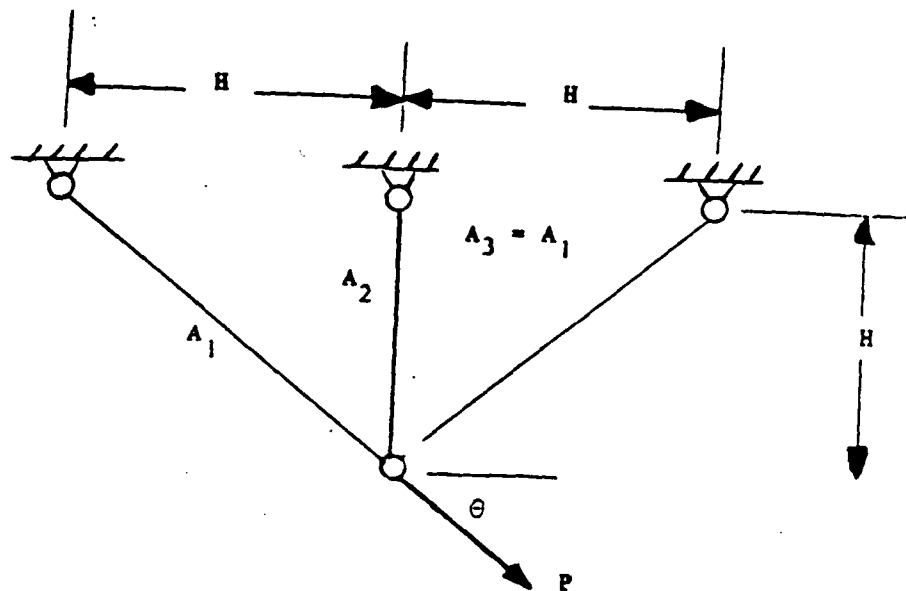


Figure 3. Three Bar Truss for Example Problem.

The Nonlinear Programming Formulation

$$\text{Minimize Volume} = 2 \sqrt{2} x_1 + x_2 \quad \bar{x} = \begin{Bmatrix} A_1, A_3 \\ A_2 \end{Bmatrix}$$

subject to

$$g_1(\bar{x}) = 20,000 - |\sigma_1| \geq 0$$

$$g_2(\bar{x}) = 20,000 - |\sigma_2| \geq 0$$

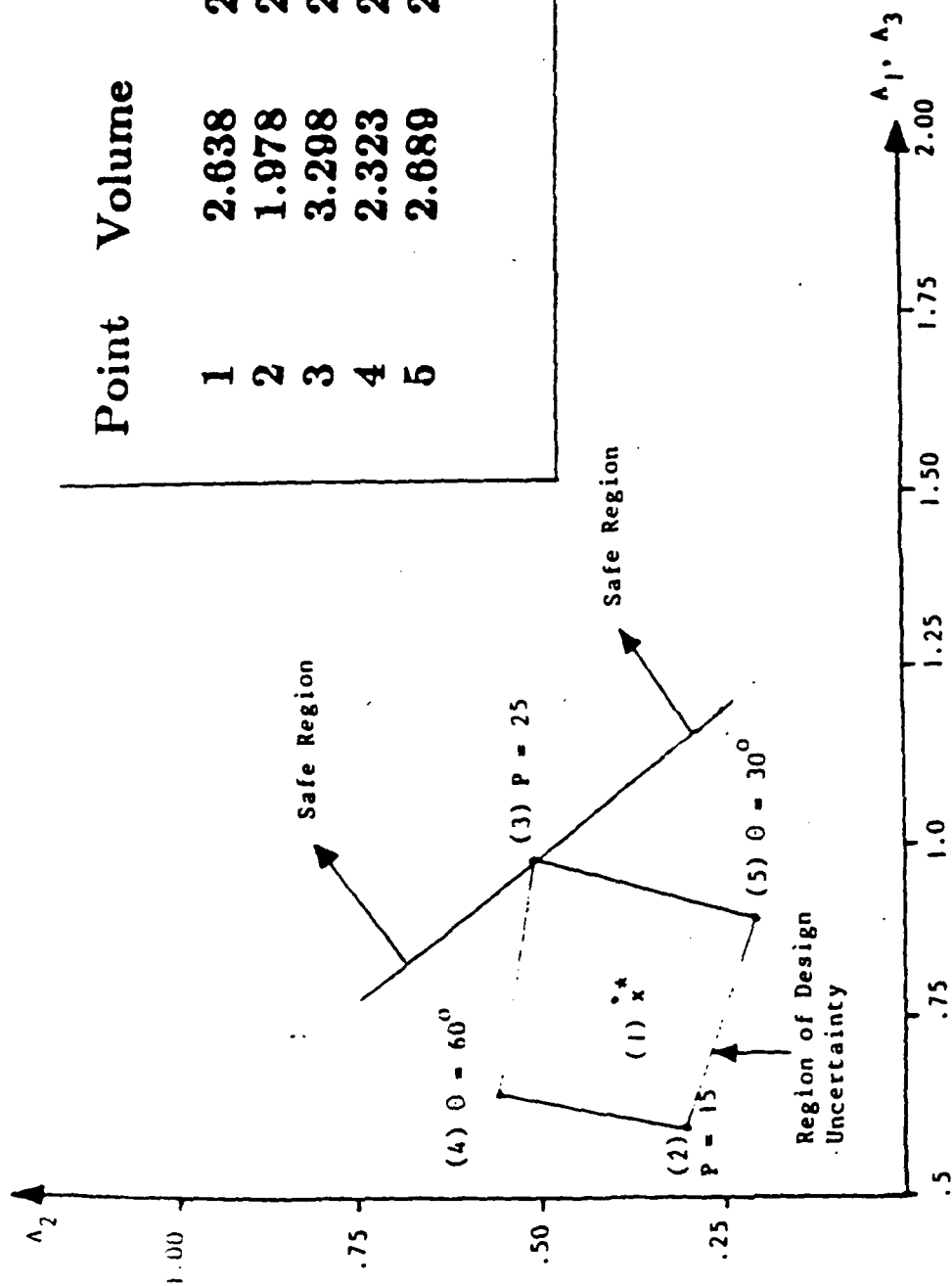
$$g_3(\bar{x}) = 20,000 - |\sigma_3| \geq 0$$

and

$$.01 \leq \bar{x}_i \leq 5.0 \quad i=1,2$$

Variations Considered in Truss Design

- Change in load from 15,000 lb to 20,000 lb
- Change in direction of load from 30 to 60 degrees



Point	Volume	σ_{\max}
1	2.638	20 ksi
2	1.978	20 ksi
3	3.298	20 ksi
4	2.323	20 ksi
5	2.689	20 ksi

Figure 4. Optimal Design for 3 Bar Truss with Variations in Load Magnitude and Direction.

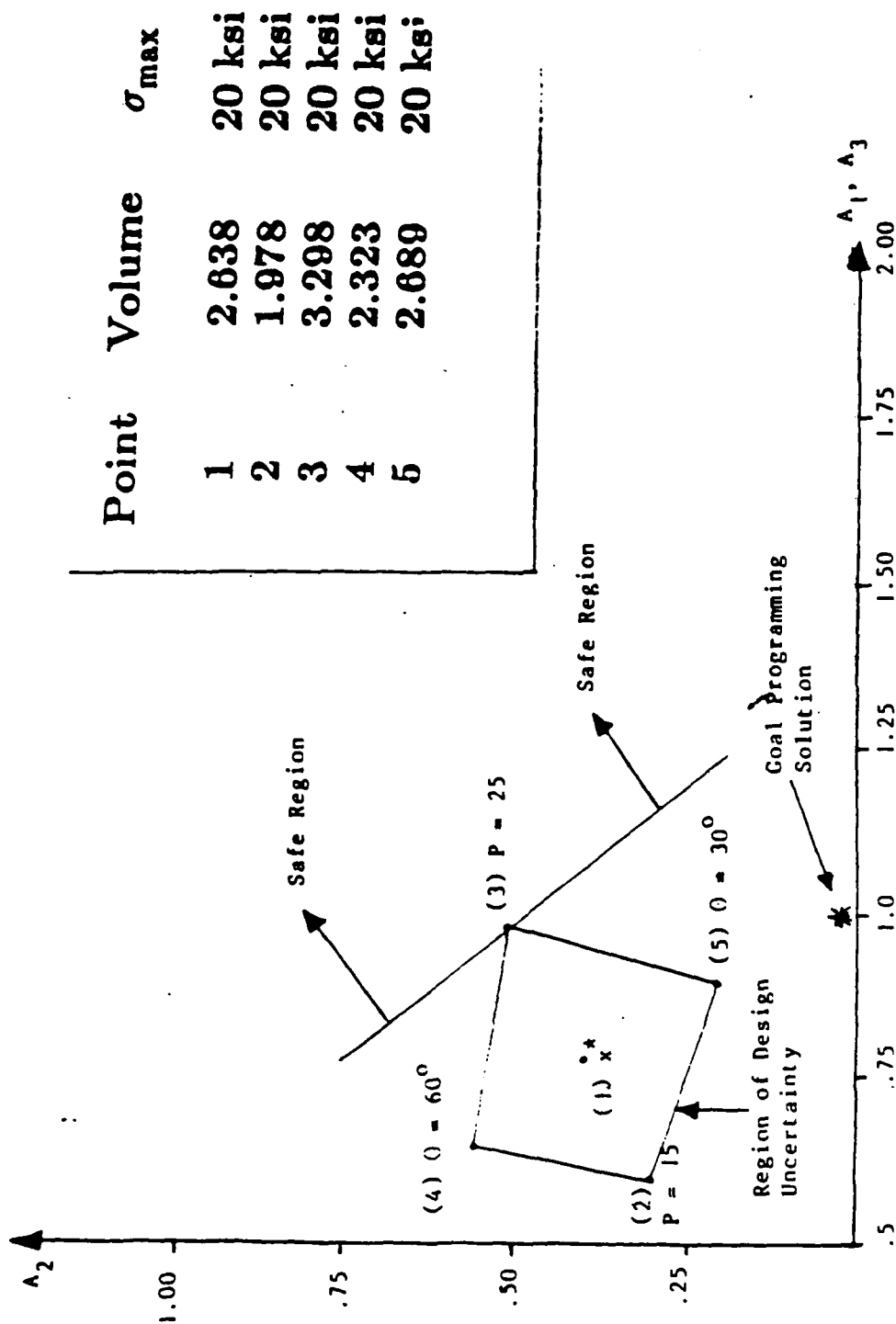


Figure 6. Optimal Design for 3 Bar Truss From
Goal Programming Formulation Number
Two.

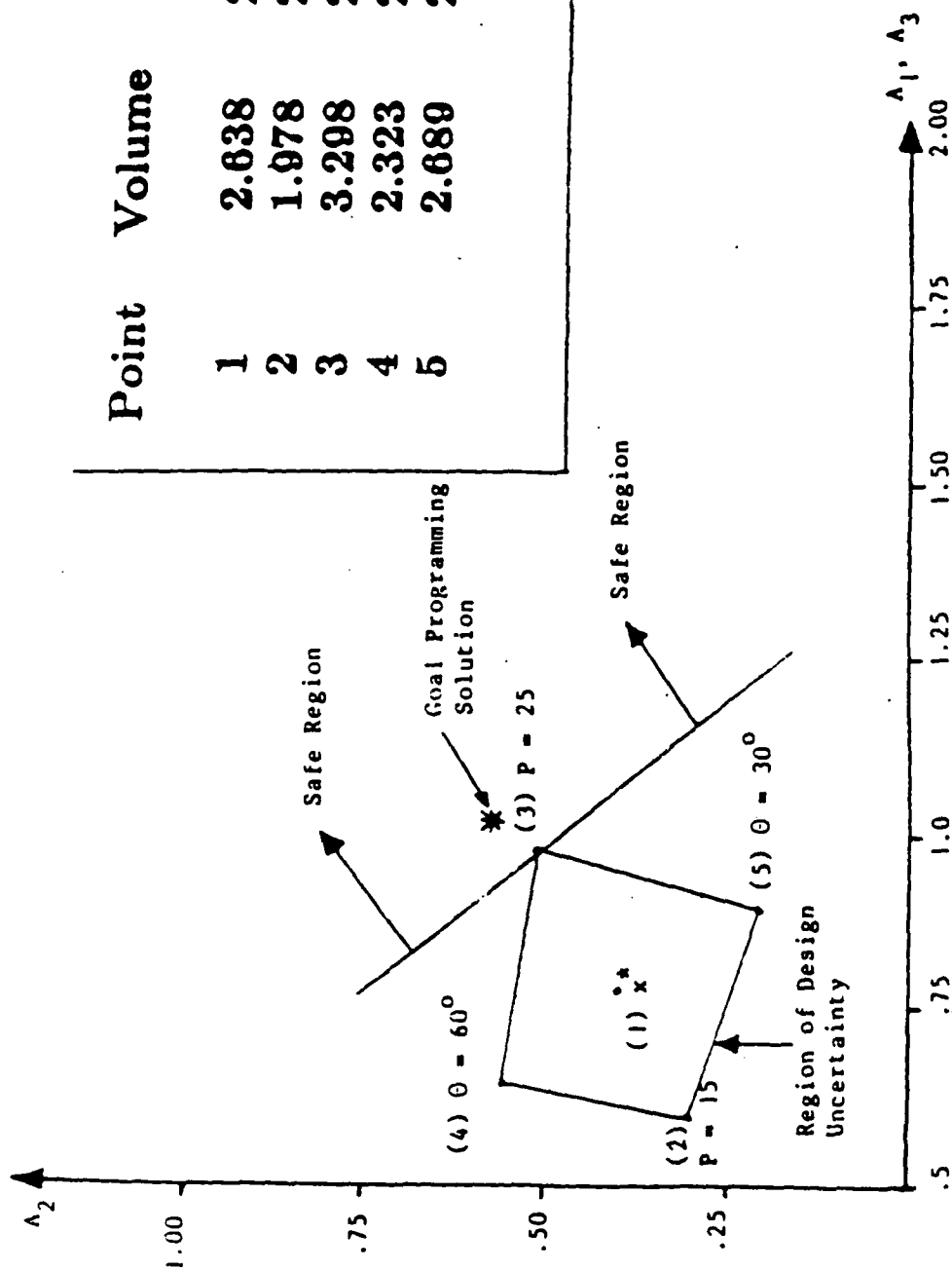


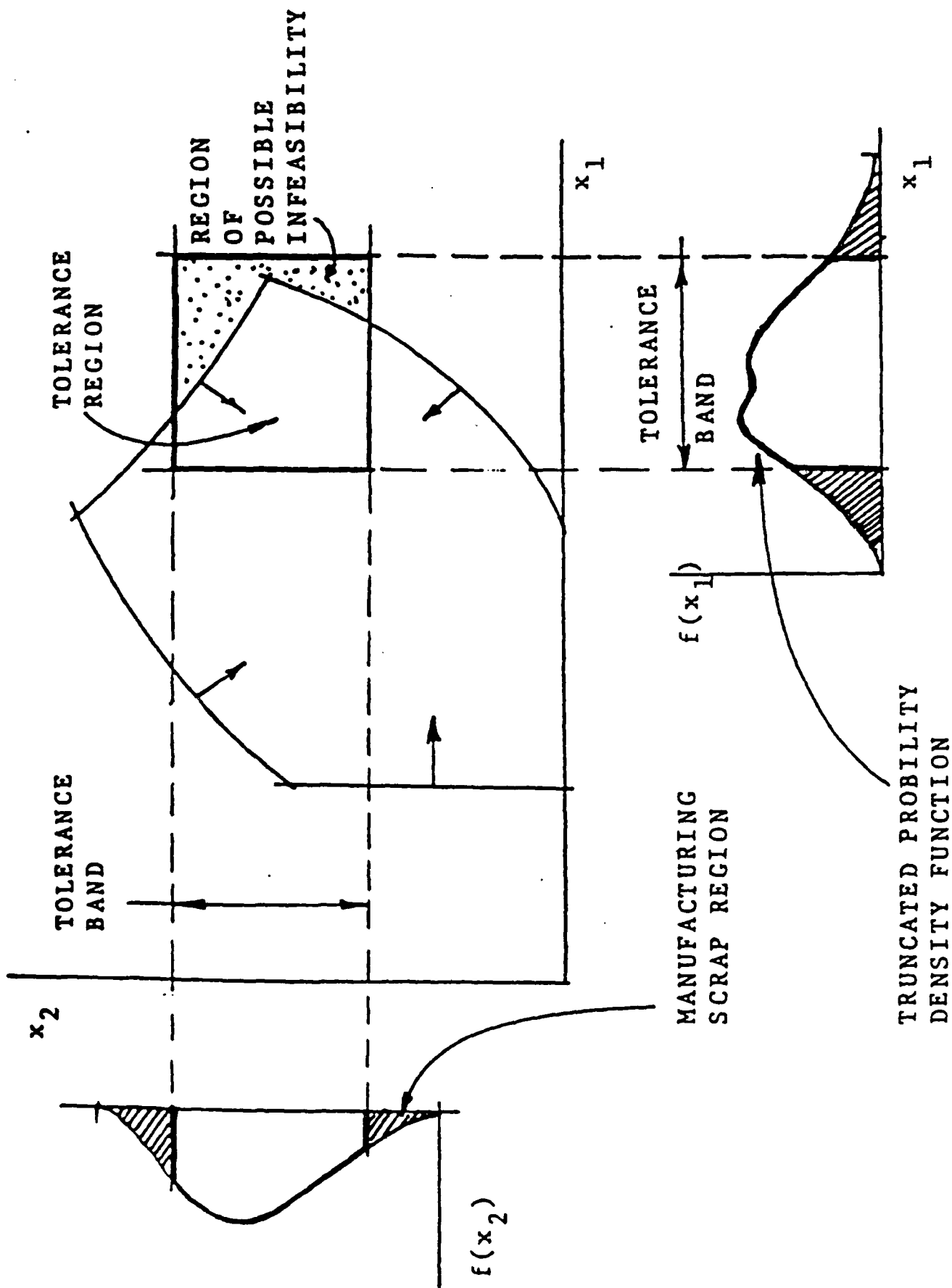
Figure 5. Optimal Design for 3 Bar Truss From
Goal Programming Formulation Number
One.

Goals for Nonlinear Goal Programming

- Minimize Volume: $\text{Volume} + d_1^- - d_1^+ = 0$
- Minimize Max Stress: $\text{Max Stress} + d_2^- - d_2^+ = 0$
- Minimize change in stress due to design change.

OBSERVATION:

If we can precisely define the uncertainties for our design then they are not really uncertain.



Session V.

Statistical Issues in
Integrated Circuit Design
and Fabrication

**STATISTICAL APPROACHES TO
IC DESIGN AND FABRICATION
PROCESS CONTROL**

Andrzej J. Strojwas

**Carnegie Mellon University
Pittsburgh, PA 15213**

**UIED-88
May 11, 1988**

PRESENTED APPROACH:

EIGHT YEAR RESEARCH PROGRAM AT CMU

MAIN CONTRIBUTORS:

S. W. DIRECTOR

W. MALY

S. R. NASSIF

C. SPANOS

P. K. MOZUMDER

OUTLINE

Stochastic nature of VLSI fabrication

Statistical simulation

Tuning - parameter identification

Worst-case analysis/design

Yield maximization

Statistical quality control

VLSI DESIGN AND MANUFACTURING

Goal:

minimize cost per chip

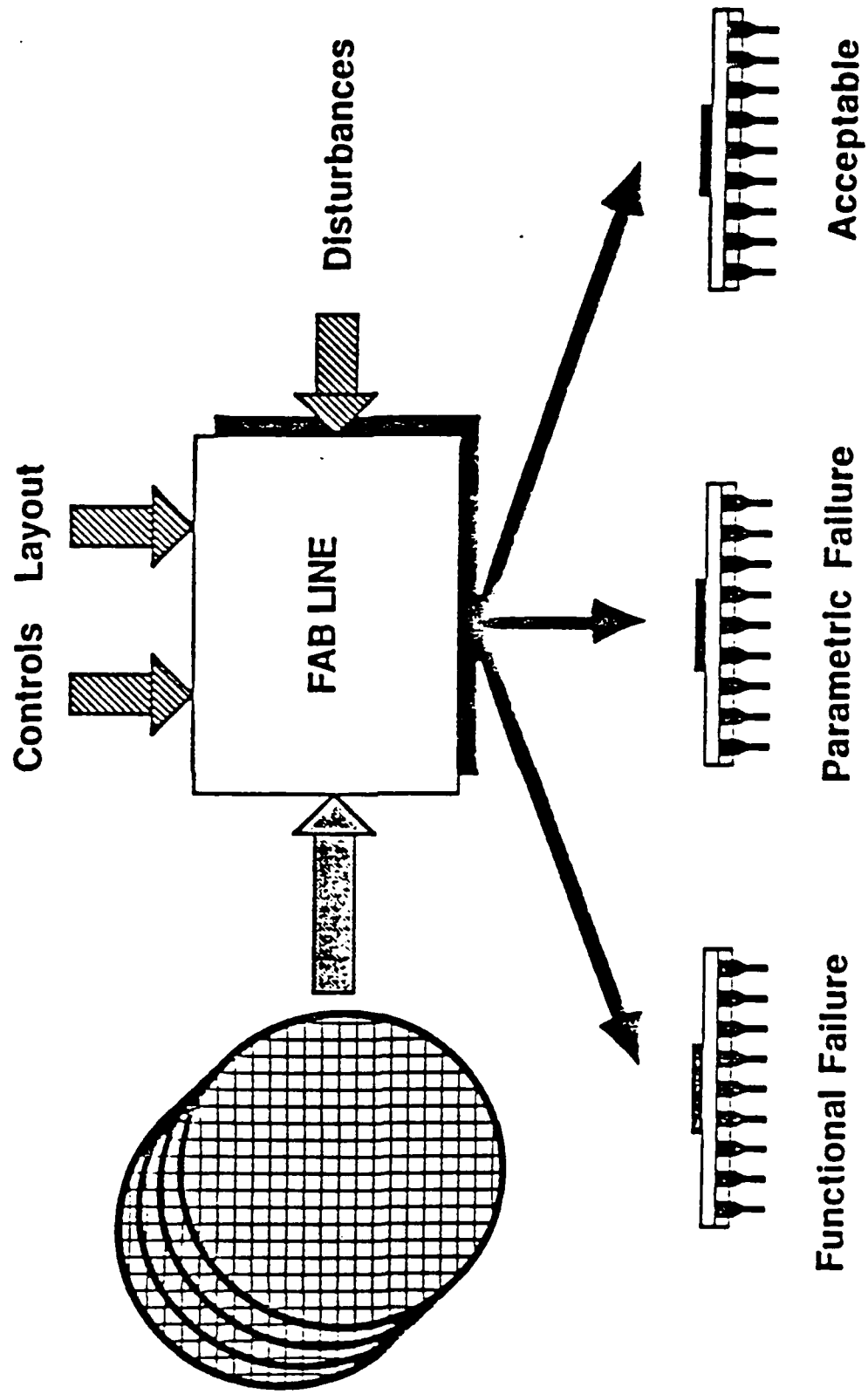
- Short design cycle
- Fast turn-around
- High manufacturing yield

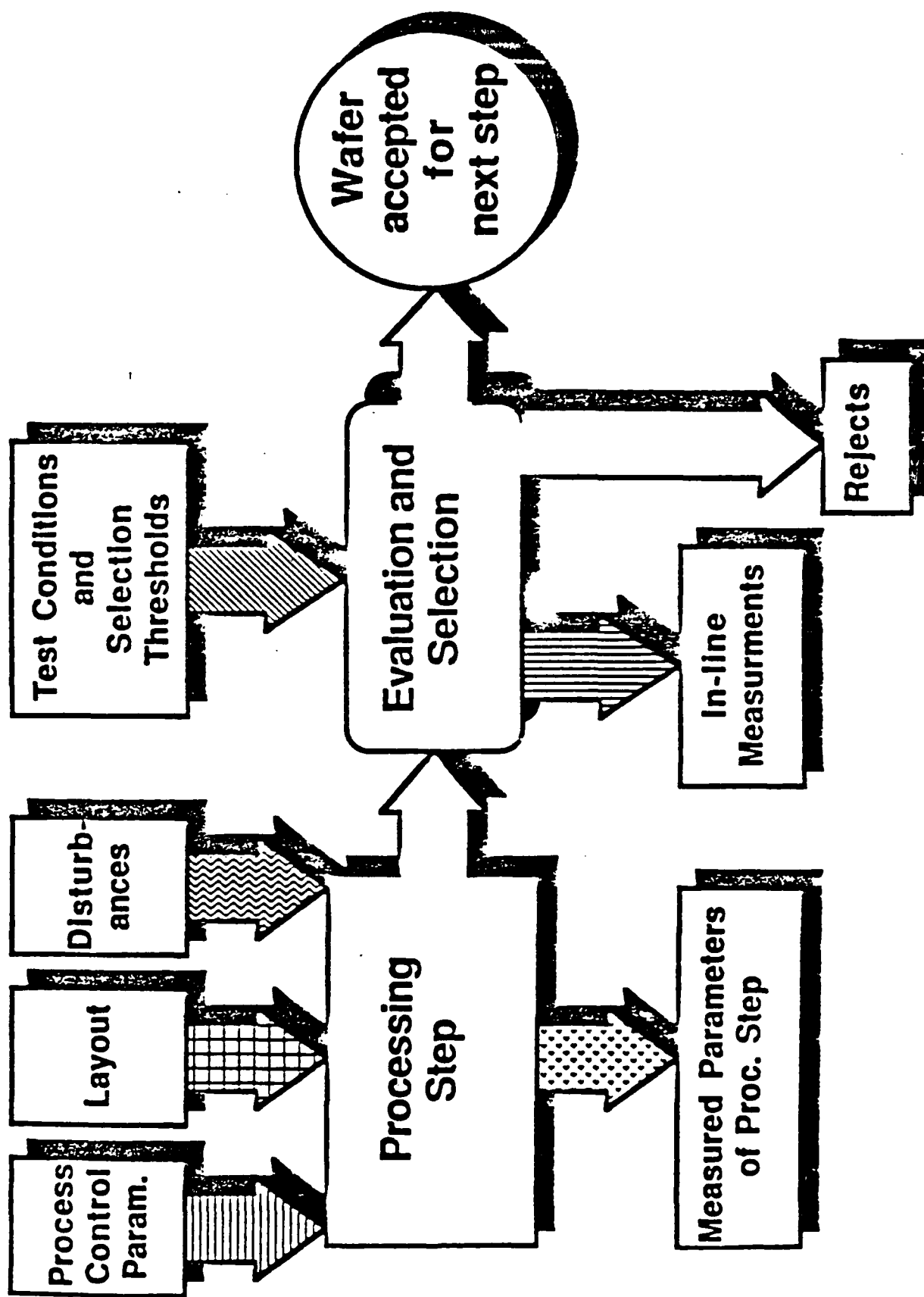
VLSI DESIGN AND MANUFACTURING

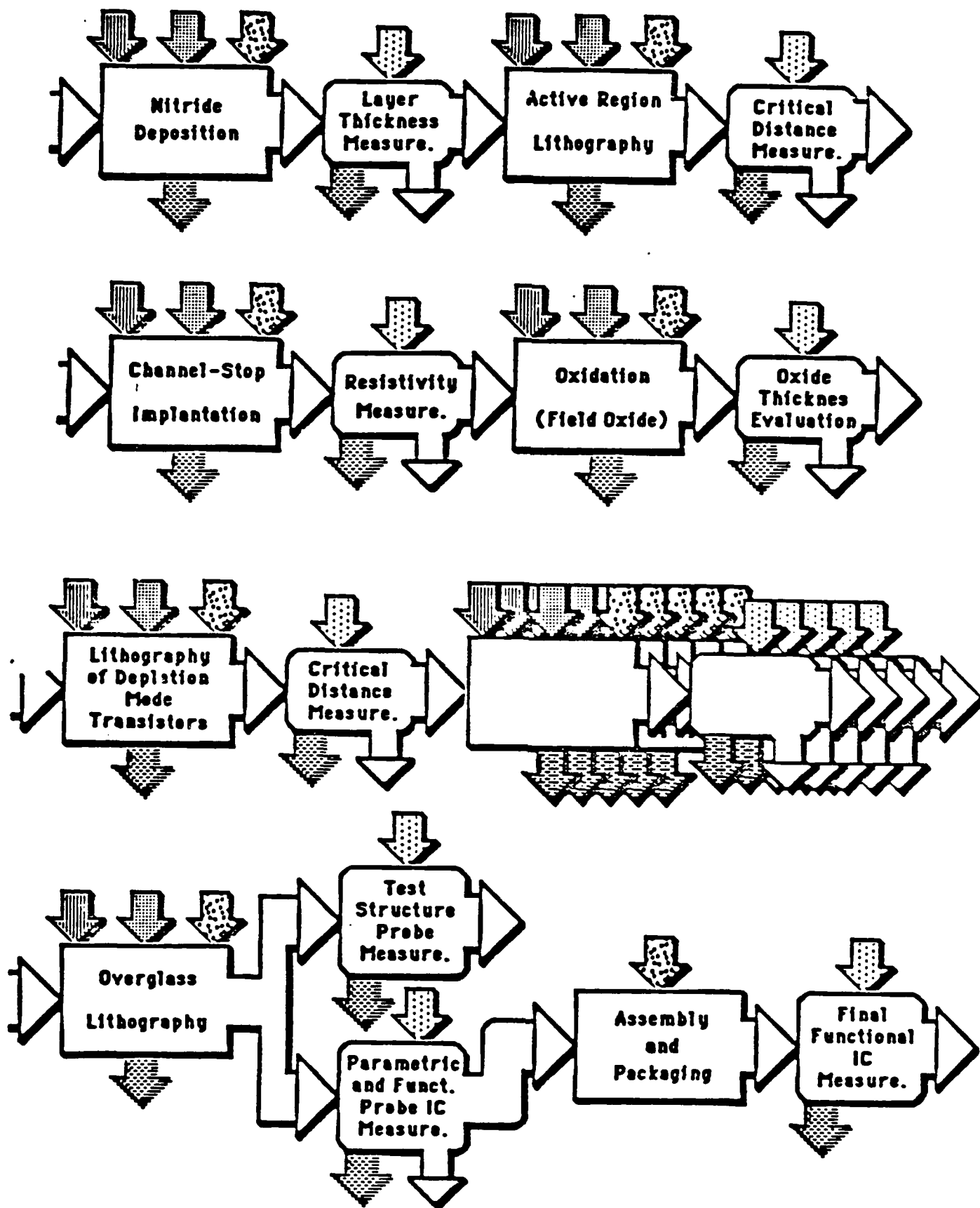
Solutions:

- Design Automation
 - system, circuit and process levels
- Yield Maximization
 - parametric and catastrophic
- *Active* Process Control
 - monitoring, diagnosis, quality and adaptive control

Typical IC Fabrication Line







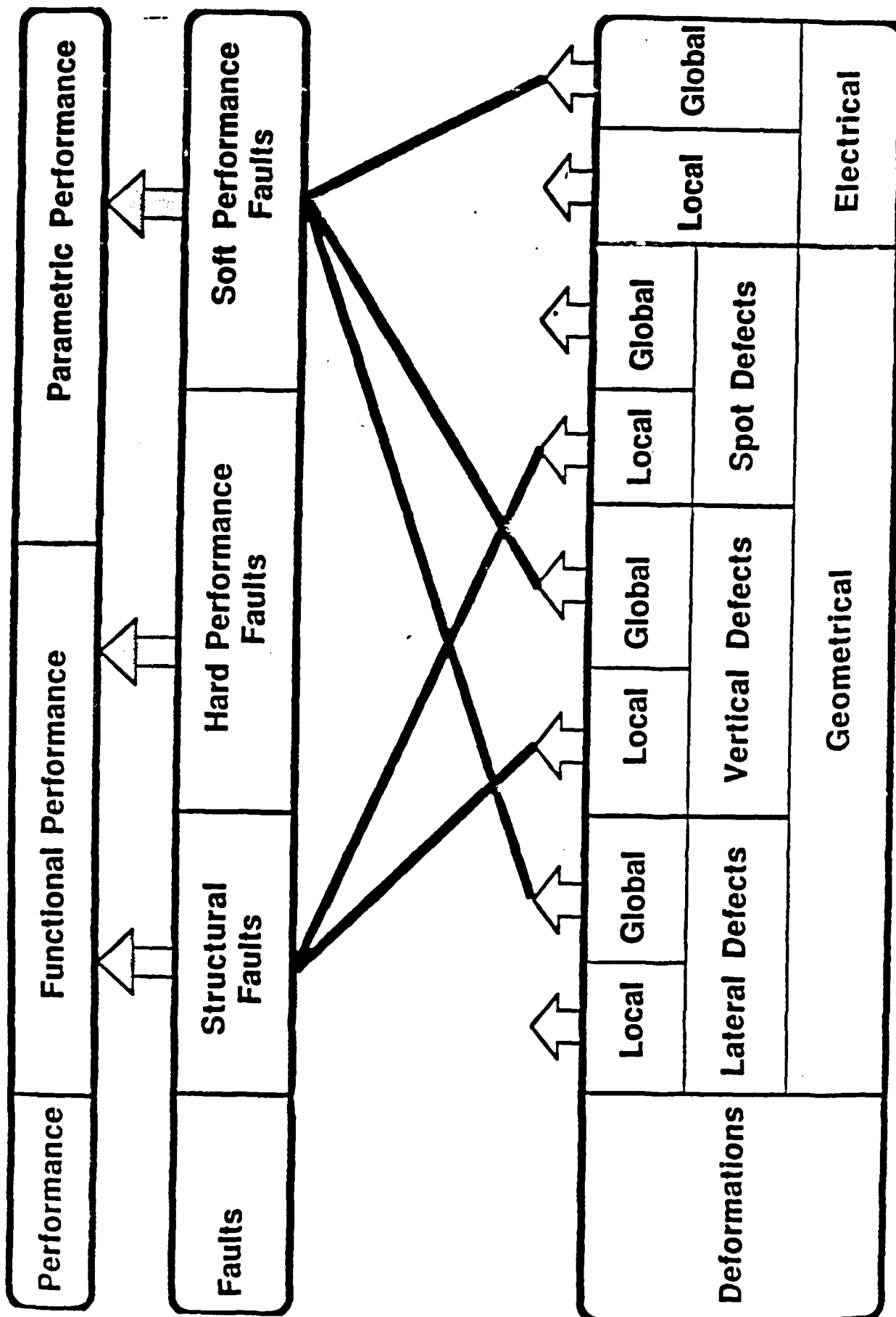
Causes of Process Disturbances

- human error and equipment failures
- fluctuations in process conditions
e.g., turbulent gas flow
- fluctuations in materials
e.g., impurities in chemicals
- variations in substrate
*e.g., point defects, dislocations
surface imperfections*
- lithographic spots (during mask fab and use)
e.g., transparent spots in opaque regions

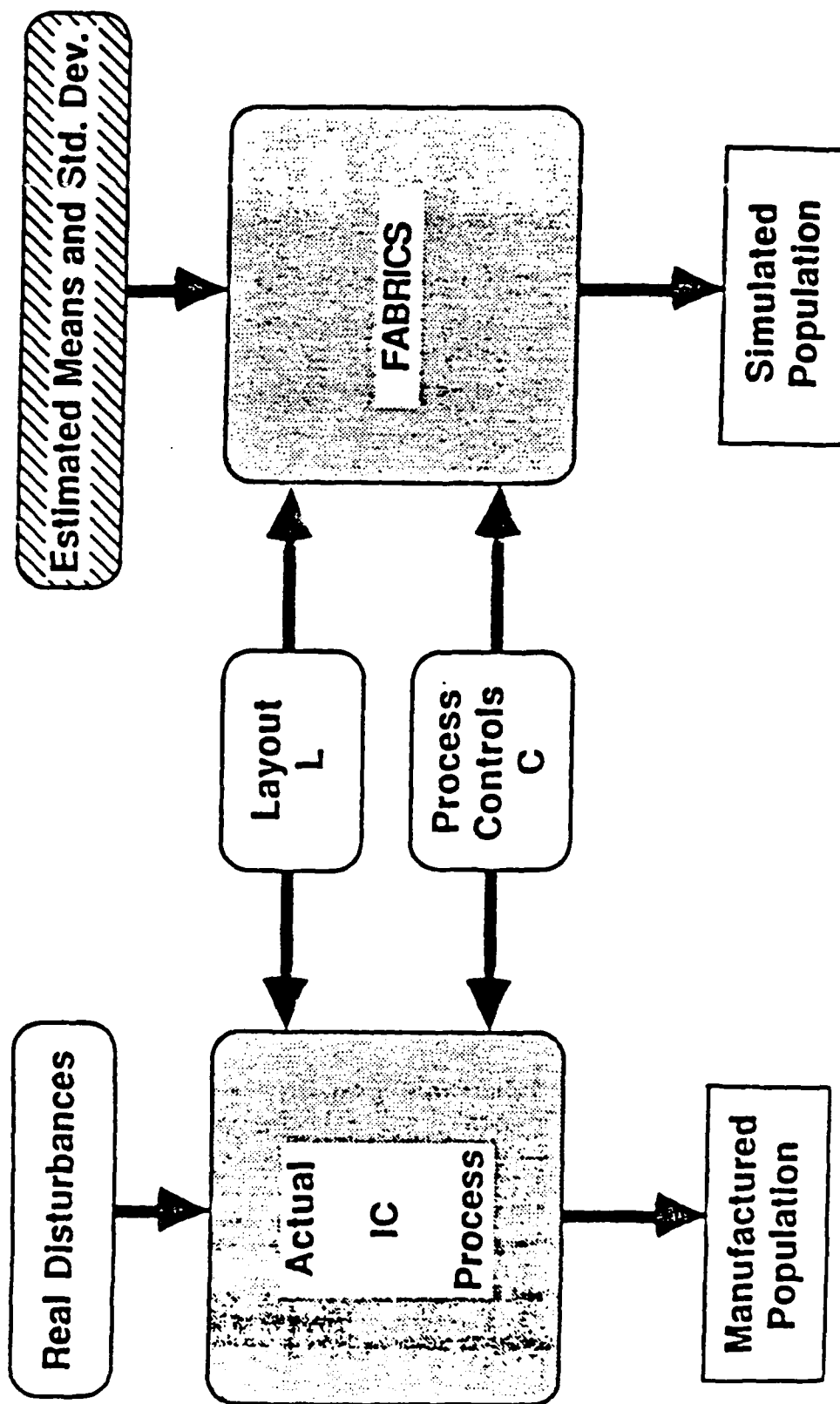
Performance Faults

- **Structural faults:**
 - Changes in circuit topology (e.g., shorts or opens)
 - may depend on bias
- **Hard performance faults**
 - IC doesn't function properly (e.g., some state transitions do not occur)
- **Soft performance faults**
 - IC functions but response (e.g., speed or power) falls outside allowable limits

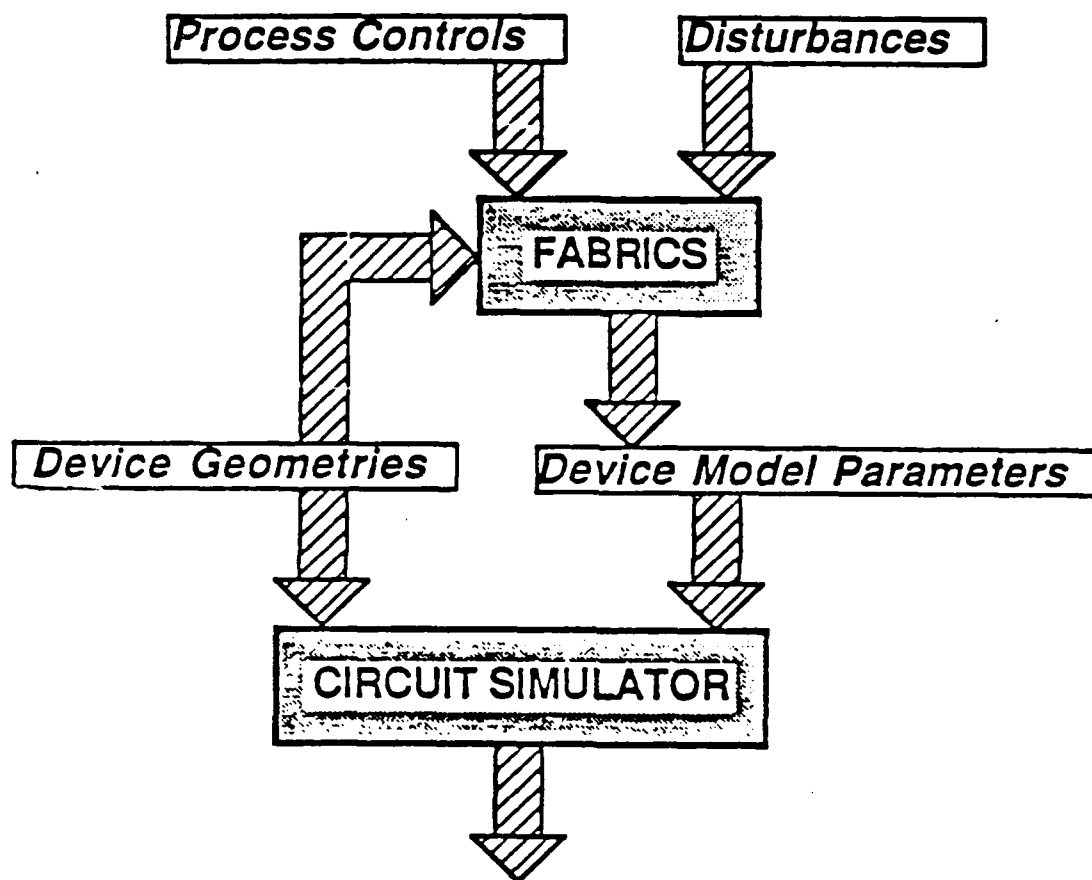
Primary Dependence of Faults on Disturbances



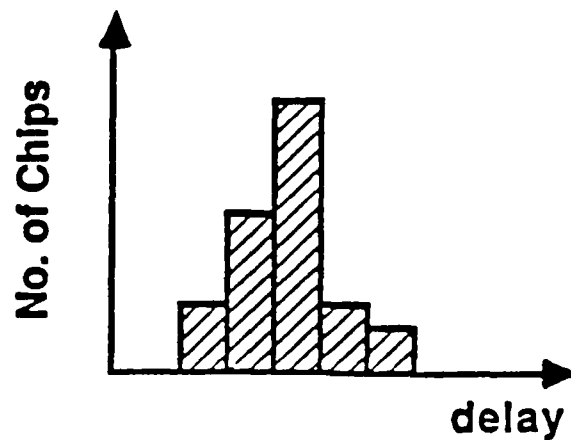
Simulation of the Fabrication Process



Manufacturing Based Simulation

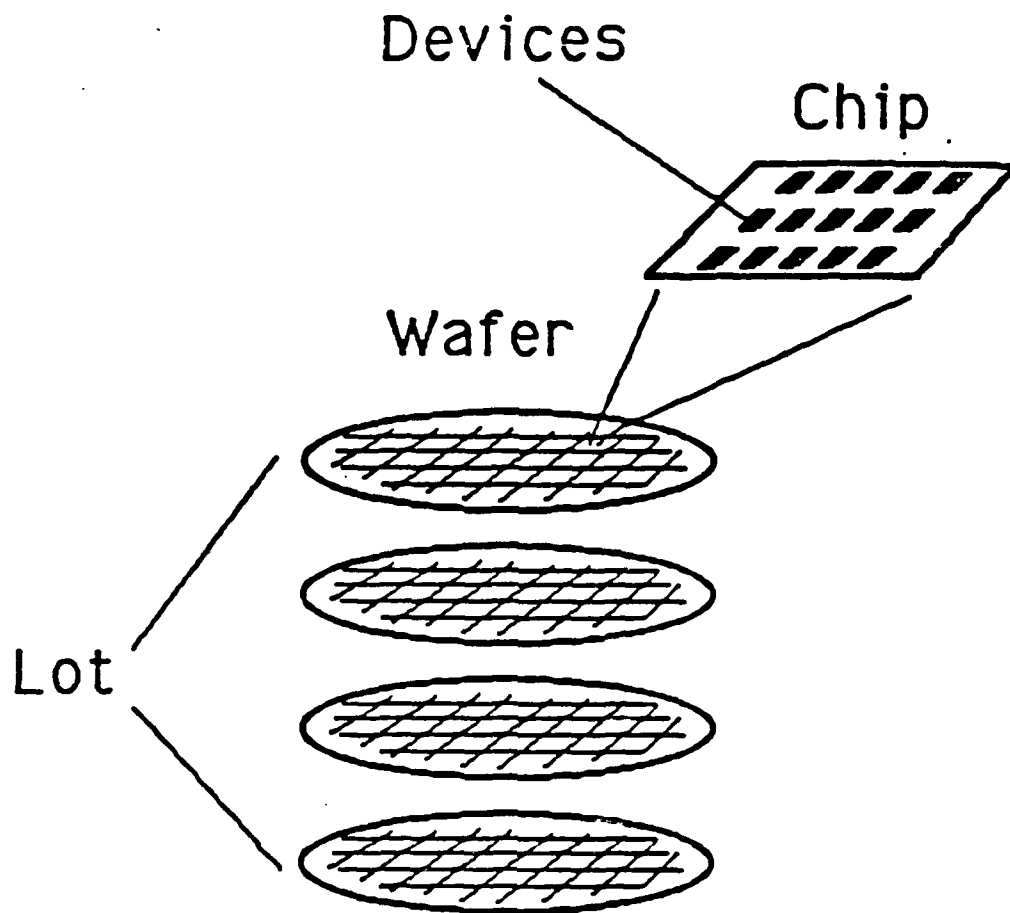


IC Performance



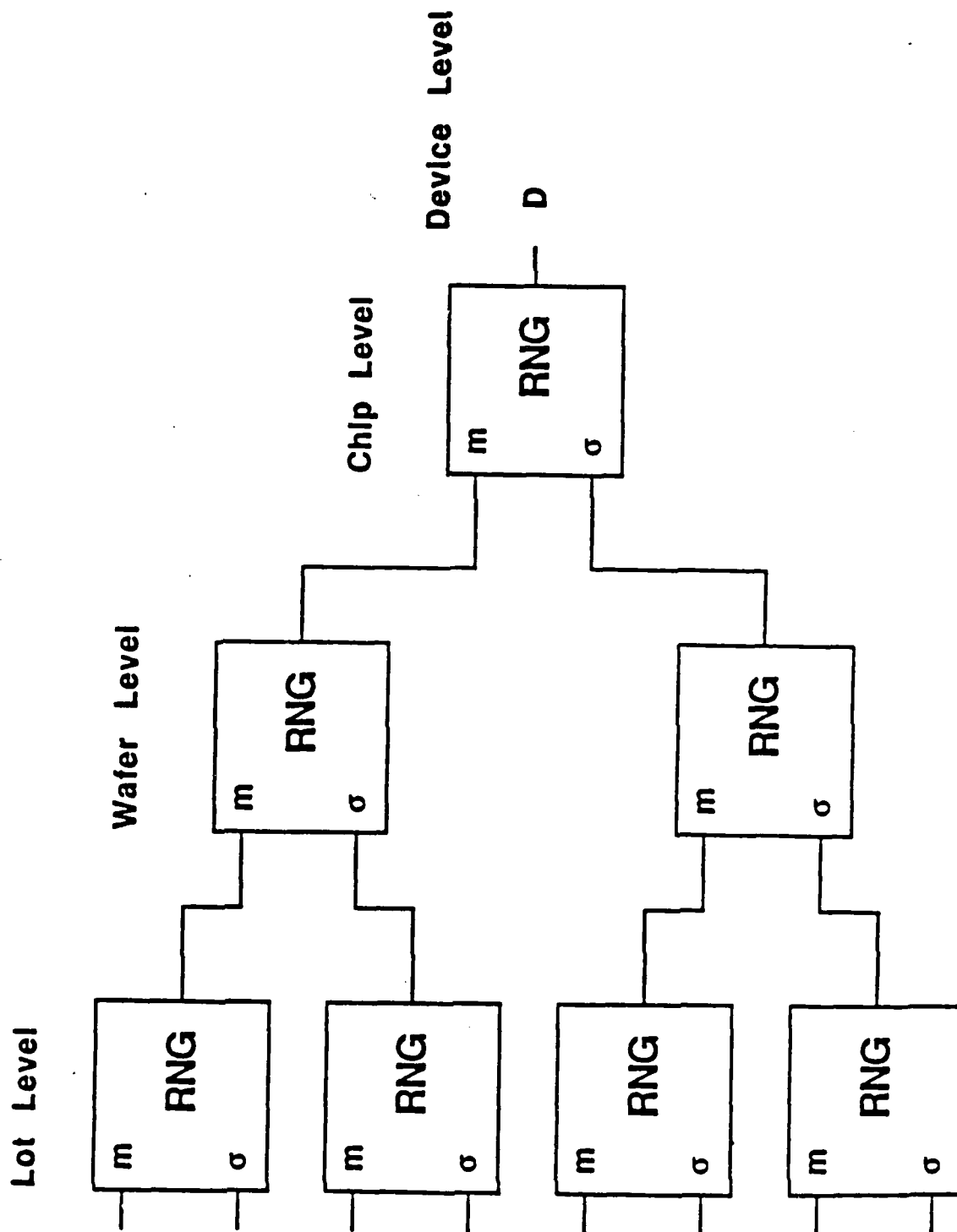
PROCESS DISTURBANCES

- **Fluctuations in process controls (e.g. temperature variations) account for small portion of variations in device performance**
- **Physical disturbances (e.g. diffusion coefficient, oxide growth rate) crucial to model device parameter fluctuations realistically**

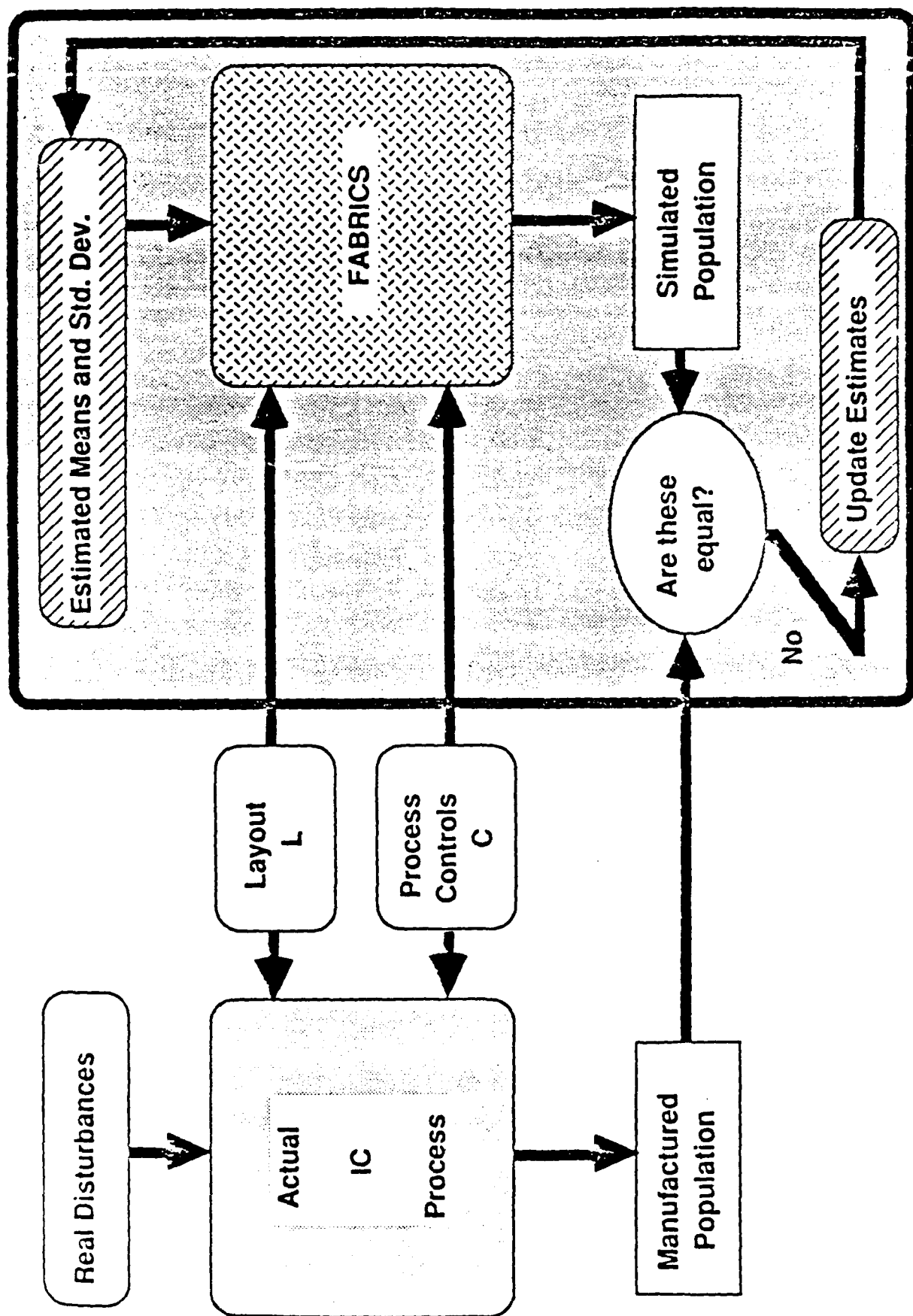


Hierarchical Nature of the IC Process

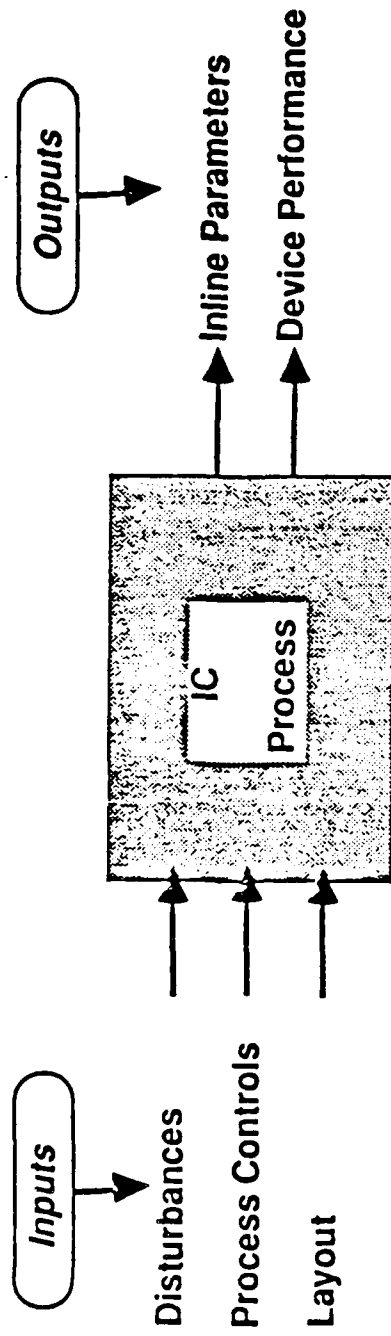
Generation of Process Disturbances



Prometheus - A Tuning Tool for FABRICS



Assumptions



- Disturbances are independent, normal and hierarchical
- Outputs have a unimodal jpdf
- Output distribution may be transformed to joint normal distribution
(i.e. log, exp., square, etc.)
- Smooth function relates inputs to transformed outputs

Concerns / Approach

Concerns:

- high dimension (40 disturbances, 4 moments per disturbance)
- hierarchy makes statistical verification difficult
- high computational cost of simulation

Approach:

- dimension reduction through sensitivity analysis
- decomposition into sequence of "flat" minimization problems
- use of regression models for simulation

Sensitivity Analysis and Decomposition

Sensitivity:

- For each output j , determine degree of good polynomial fit for each disturbance i (depends on some threshold)

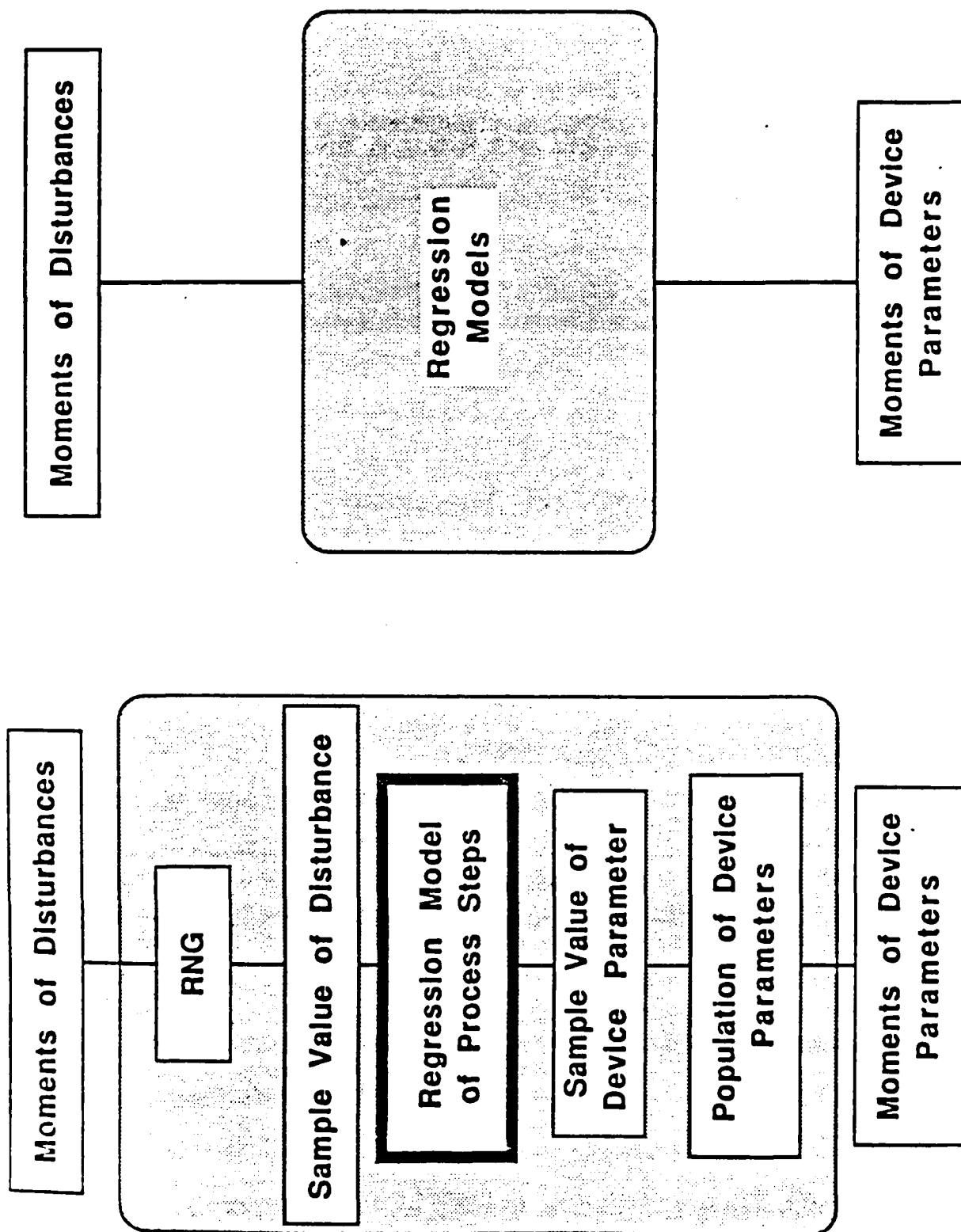
$$y^j = p_{ij} (d^i)$$

- Degree of above polynomial defines dependency factor
- Selection of points for polynomial fitting
 - select typical device
 - choose n dimensional region about this point (n disturbances)
 - select points along each disturbance axis according to Chebyshev interpolation formula
 - simulate using FABRICS at each point

Decomposition:

- Use sparse ordering matrix methods to decompose

Regression Models of Moments



Decomposed Flat Statistical Tuning

Notation:

Θ_k^y = vector of moments of for output jpdf in subproblem k

Θ_k^d = vector of moments disturbance jpdf in subproblem k

Problem Formulation:

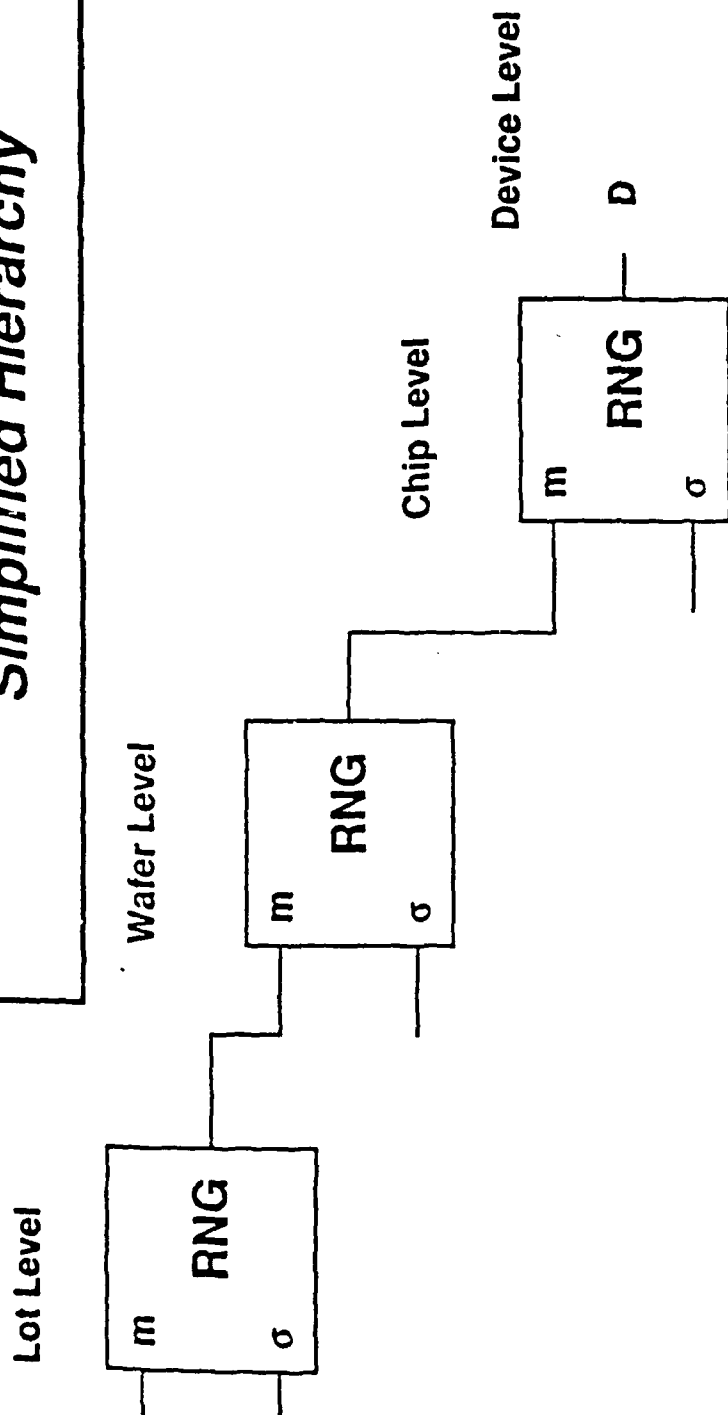
$$\min_{\Theta_k^d} || \mathbf{w}_k^T [\tilde{\Theta}_k^y - \tilde{\Theta}_k^d] ||_2$$

where

$$\tilde{\Theta}_k^y = \tilde{\mathbf{G}}_k (\Theta_k^d)$$

subject to Θ_k^d lying in region defining approximation

Simplified Hierarchy



value of k^{th} process disturbance for i^{th} chip and j^{th} device:

$$d_{ij}^k = \mu_k^d + w_{ki}^d + c_{kij}^d \quad \text{for } k = 1, 2, \dots, n_p; i = 1, 2, \dots, n_c \text{ and } j = 1, 2, \dots, n_d$$

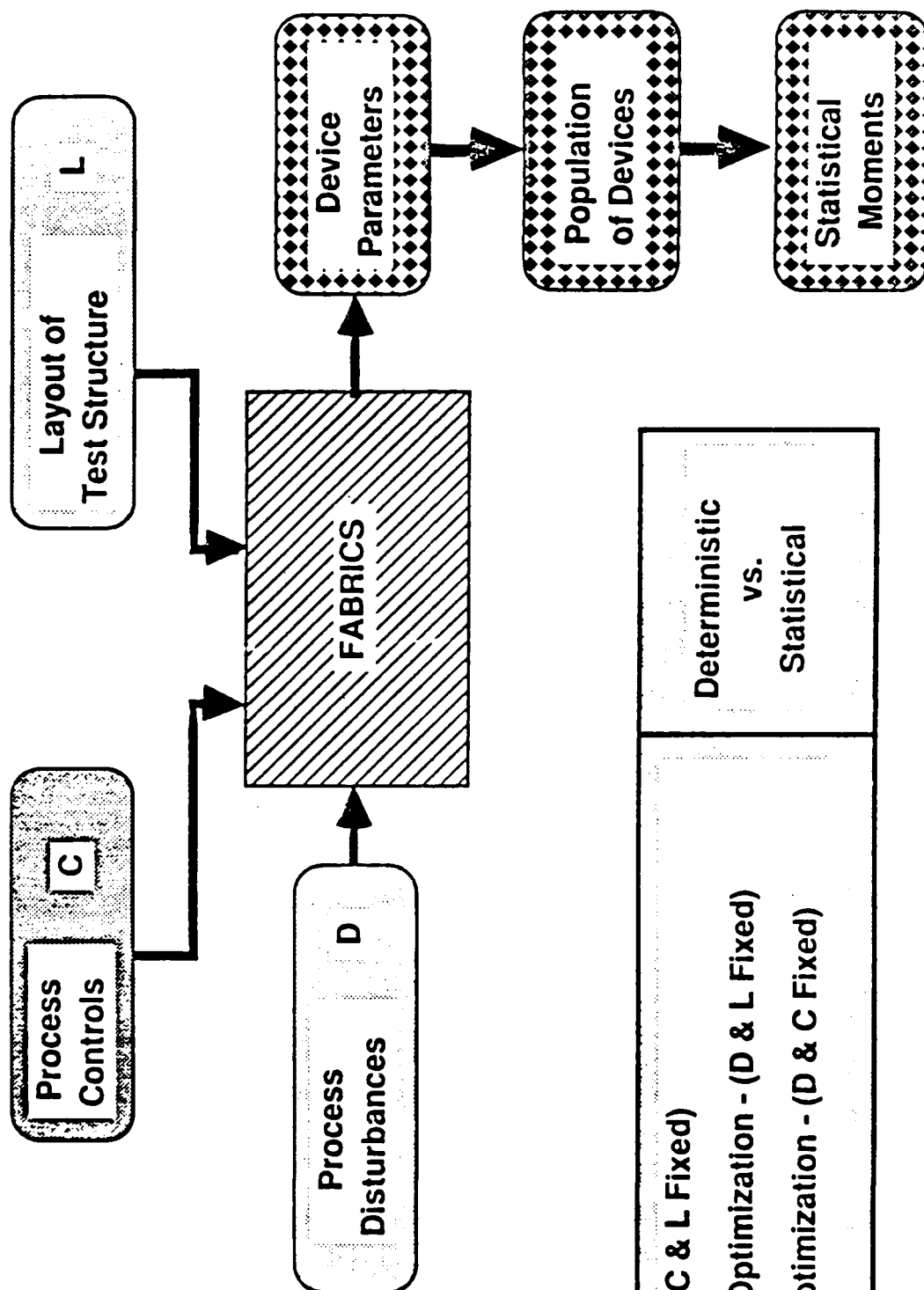
where

μ_k^d is the global wafer mean for the k^{th} disturbance

$w_{ki}^d \approx N(0, (\omega_{ki}^d)^2)$, accounts for wafer variation

$c_{kij}^d \approx N(0, (\sigma_{kij}^d)^2)$, accounts for chip variation

Tuning versus Optimization



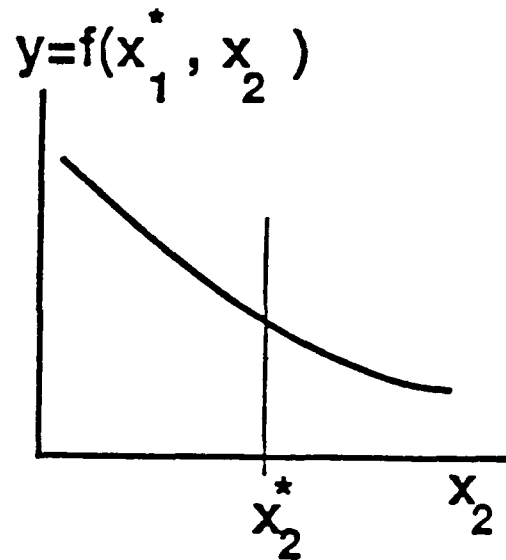
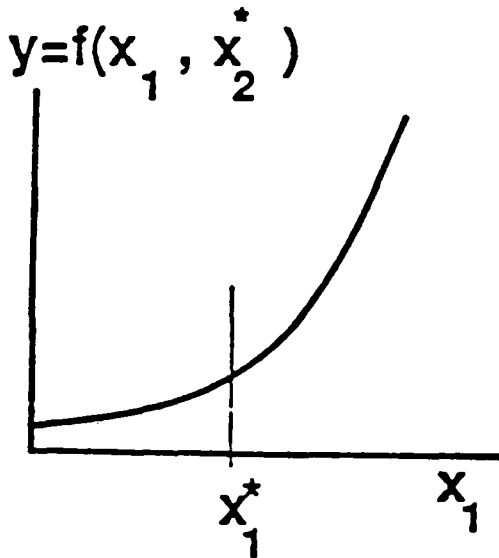
- Tuning - (C & L Fixed)
- Process Optimization - (D & L Fixed)
- Device Optimization - (D & C Fixed)

Deterministic
vs.
Statistical

Worst Case Analysis

Defn: The set of parameters which cause the performance to vary in the same direction

$$y = f(x_1, x_2)$$



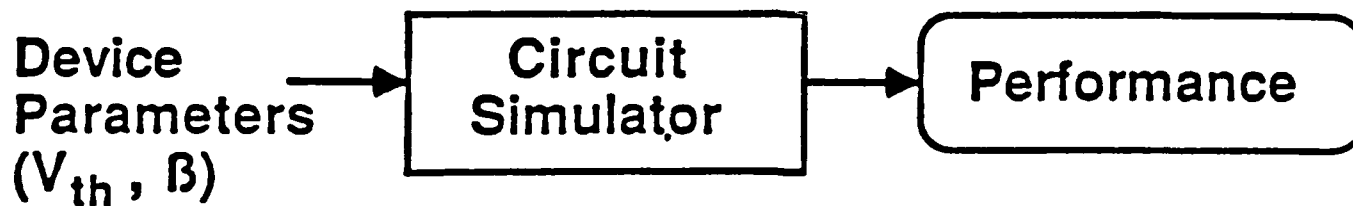
Worst Cases:

$$y \rightarrow y_{\min} : x_1^* - x_1, x_2^* + x_2$$

$$y \rightarrow y_{\max} : x_1^* + x_1, x_2^* - x_2$$

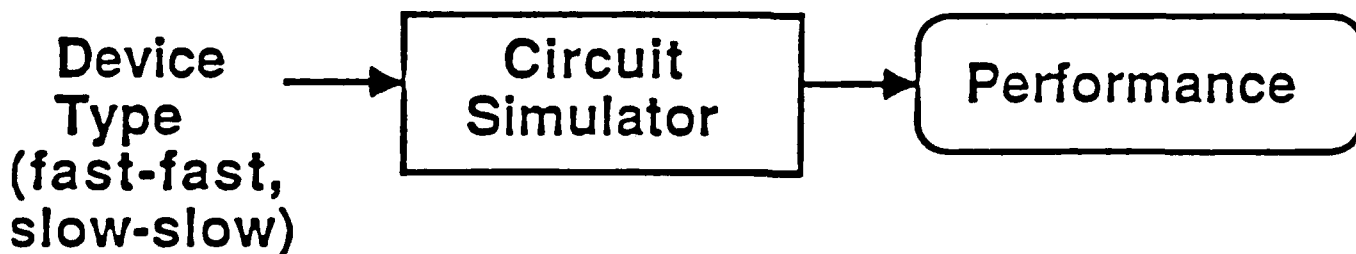
Traditional Worst Case Analysis

Parameter Based:



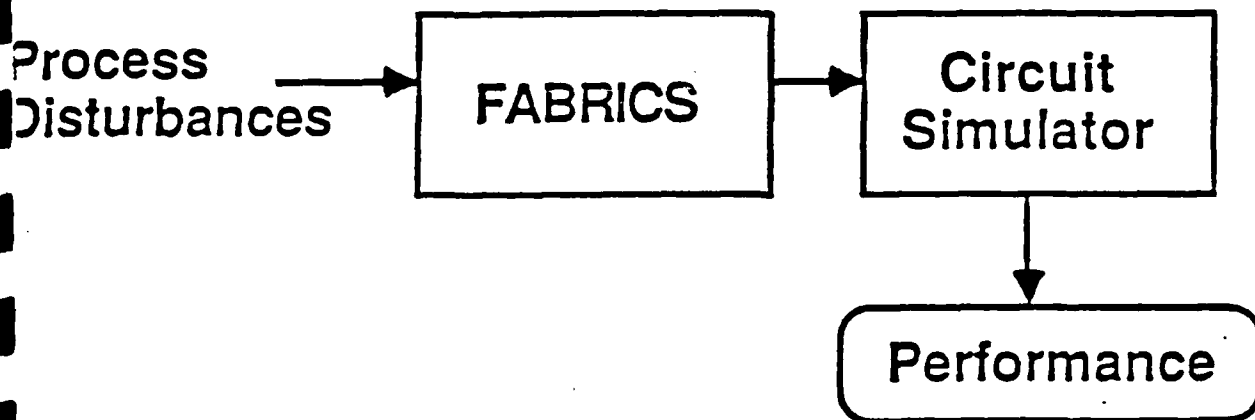
- Ignores correlation between parameters
- A set of physical parameters that produces the worst case may not exist
- Results very pessimistic

Device Based:



- Ignores correlation between devices

New Worst Case



- Process Disturbances are uncorrelated
- All device parameters and devices correctly correlated

Steps in Worst-Case Analysis

Simulate nominal performance

For each Disturbance D_j :

perturb D_j

for each performance P_i :

calculate sensitivity of P_i to D_j

determine worst case direction of D_j

For each performance P_i :

adjust disturbances in worst-case direction

simulate worst case performance

Sensitivity Analysis

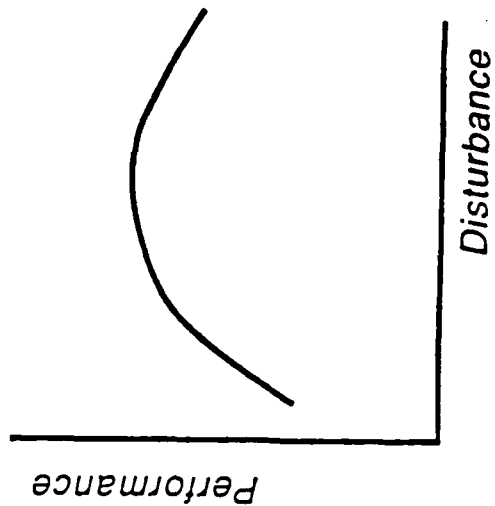
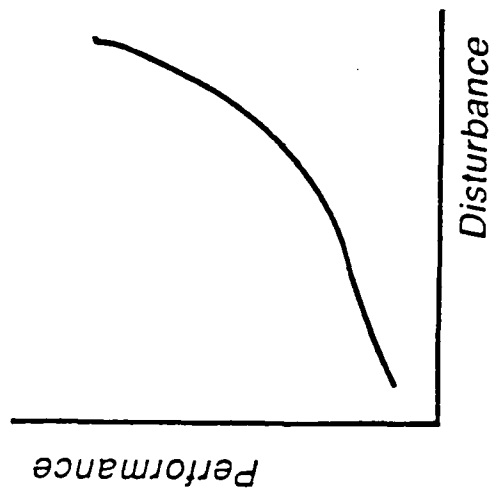
Normalized sensitivity of i^{th} performance wrt j^{th} disturbance

$$S_{ij} = \frac{\Delta P_i D_j}{\Delta D_j P_i}$$

Used to Determine:

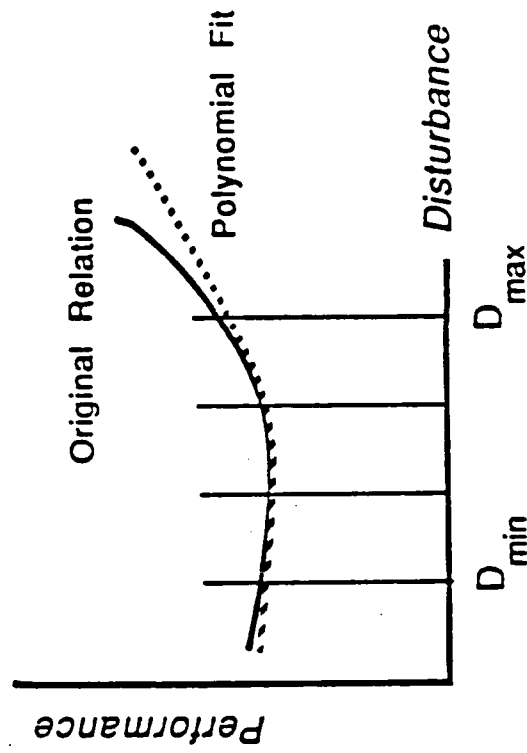
- order of magnitude of effect on disturbance
- worst-case direction

Local estimates of sensitivity may be inaccurate.



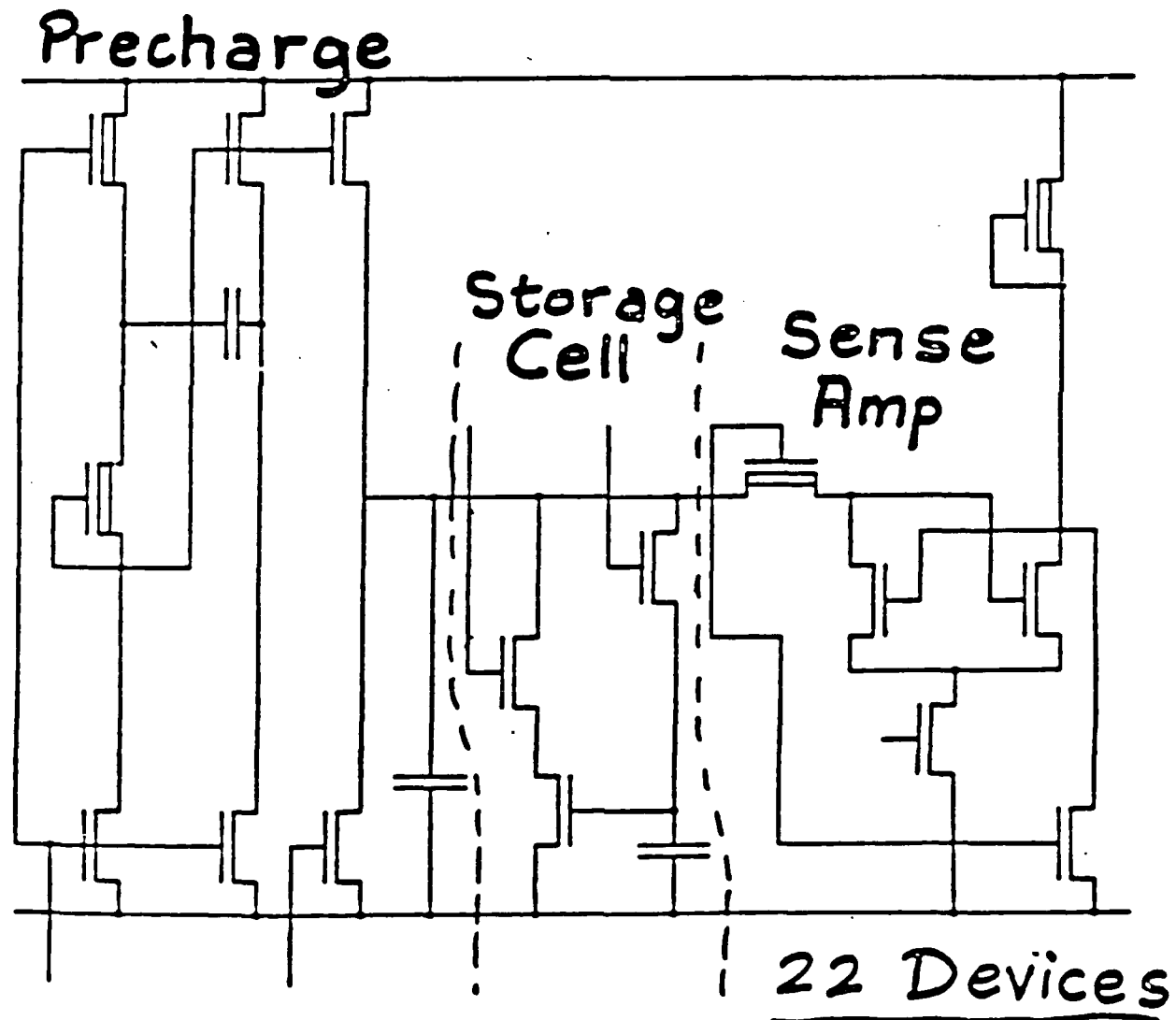
Estimating Dependence of Performance on Disturbances

- Evaluate performance over a wide range of disturbances
- Determine subset of most important disturbances
- Establish regions of monotonicity
- Build regression models for nonmonotonic relations and establish regions of monotonicity



Example: RAM

The methodology was applied to a 3-transistor cell RAM. The circuit simulated consisted of the *sense-amplifier*, the *storage cell*, the *bit-line precharge logic*, and the *read/write logic*:



Example: RAM cont.

Performances considered were:

- P, I_{peak} : power dissipation and peak current
- $\tau_{\text{read}}, \tau_{\text{write}}$: read and write times

Performances were most sensitive to:

- L_N , line width variation of nitride layer
- L_P , line width variation of poly layer
- D_B , diffusivity of Boron
- D_{As} , diffusivity of Arsenic
- R_{ox} , oxide growth rate
- N_{sub} , substrate concentration

Example: RAM cont.

The sensitivities were calculated by perturbation:

		L_N	L_P	D_B	D_{As}	R_{ox}	N_{sub}
A	P	+	-	-	-	-	-
	I_{peak}	+	-	-	-	-	-
B	τ_r	-	+	+	-	+	+
	τ_w	-	+	+	-	+	+

Two worst case parameter sets were generated.

Example: RAM cont.

Worst case performances, at $\mu \pm 1.5\sigma$:

performance	nominal	worst case A	worst case B
Power	1.45 mW	2.28 mW	1.17 mW
I_{peak}	1.1 mA	1.8 mA	0.85 mA
τ_{read}	16.9 ns	13.6 ns	20.8 ns
τ_{write}	23.8 ns	15.7 ns	28.6 ns

Probability of a performance better than this worst case
was = 0.66

Comparison to Traditional Worst Case

Performed worst case analysis with respect to device parameters.

Sensitivities were calculated by perturbation.

Performances were most sensitive to:

- Depletion device: V_{th} , KP, and γ
- Enhancement device: V_{th} , KP, and γ

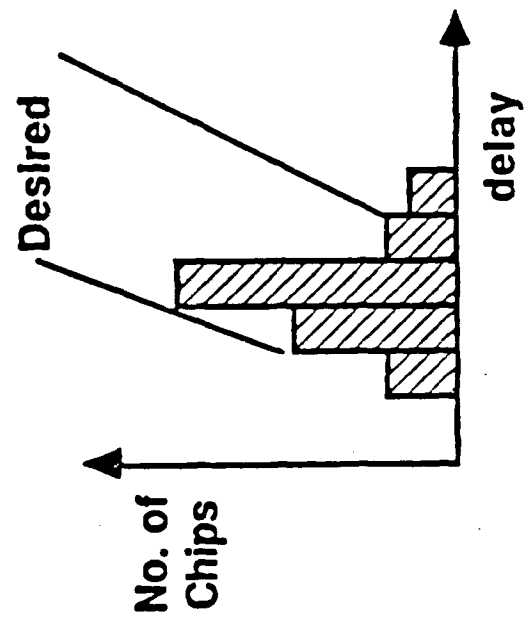
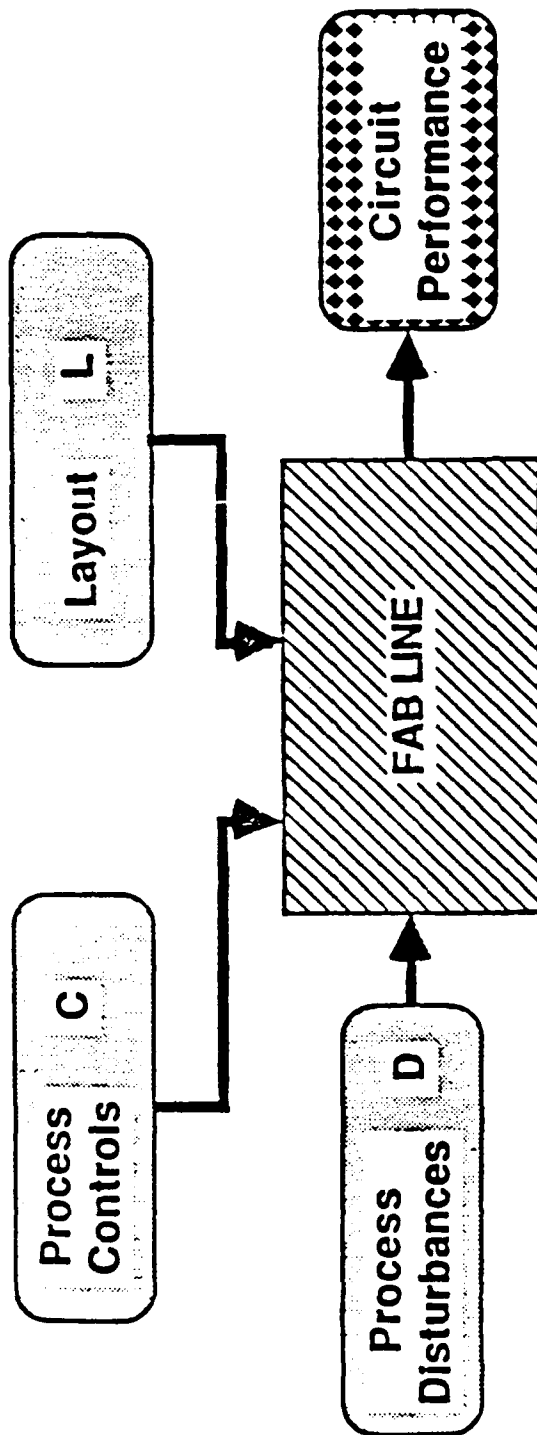
Worst case was simulated with device parameters perturbed by 1σ from their mean values.

Comparison, cont.

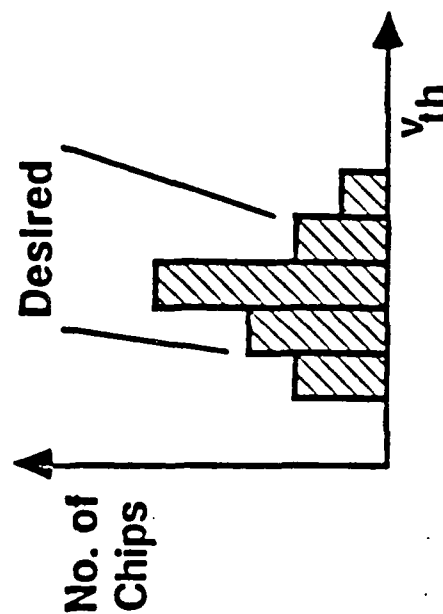
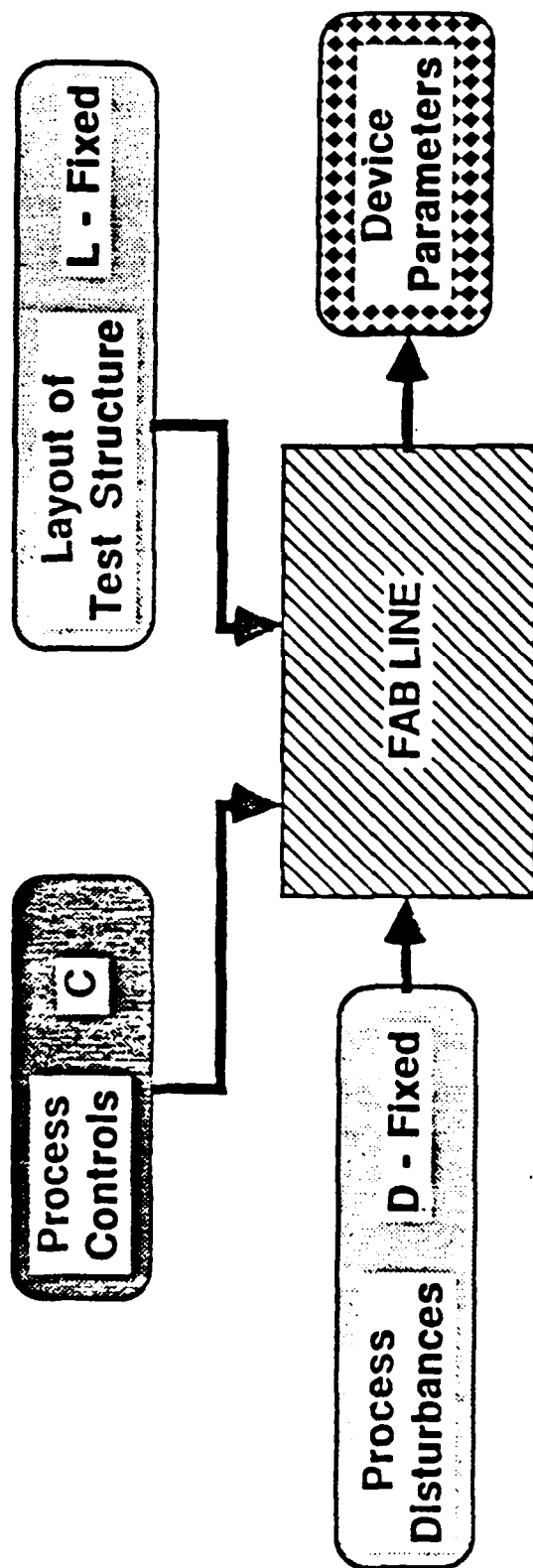
Worst cases were found at 1σ

performance	nominal	worst case Disturbances	worst case Devices
Power	1.45 mW	1.77 mW	2.30 mW
I_{peak}	1.1 mA	1.4 mA	1.9 mA
τ_{write}	16.9 ns	19.8 ns	27.0 ns
τ_{read}	23.8 ns	27.3 ns	31.0 ns

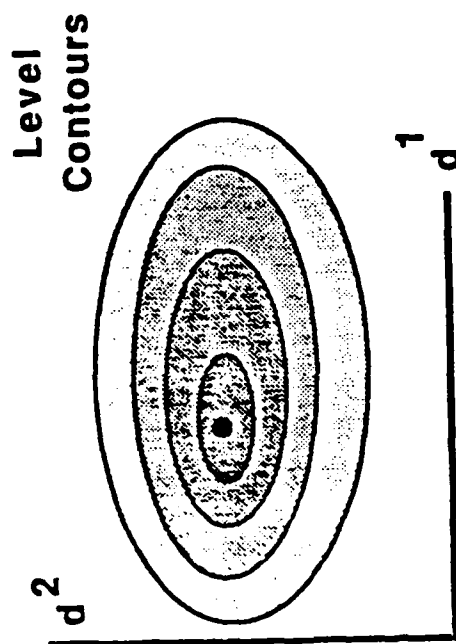
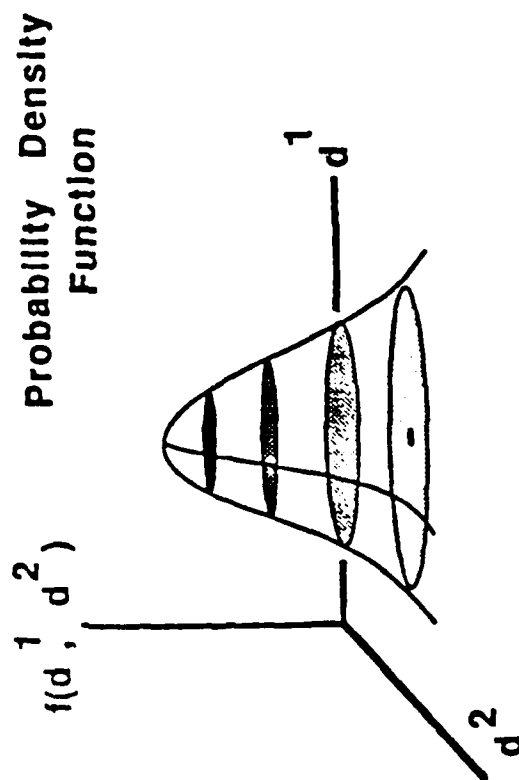
Yield Optimization



Process Optimization

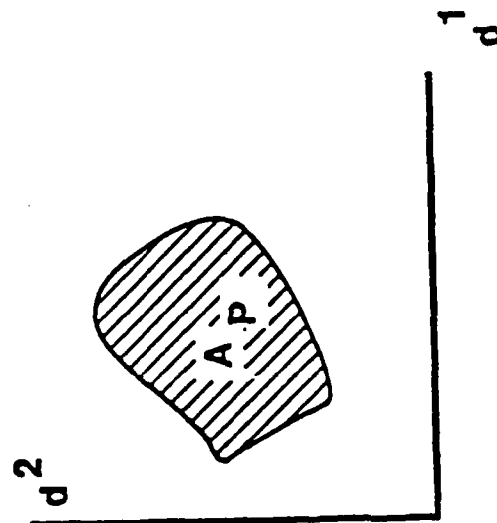


Disturbances and Acceptability Region

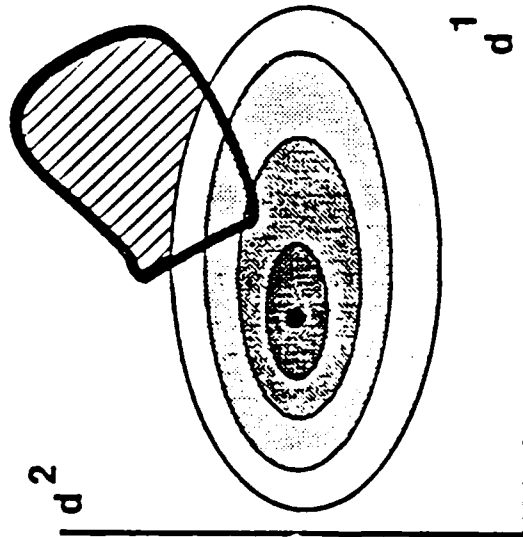


Parametric Acceptability Region:

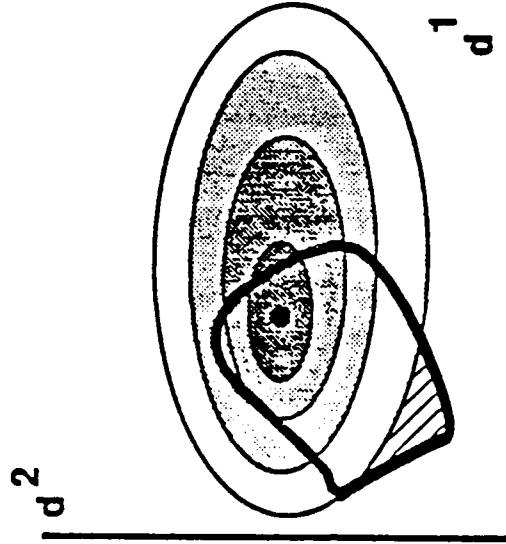
$$A_P(C, L) = \{D \mid \text{that do not disturb parametric performance}\}$$



Design Centering



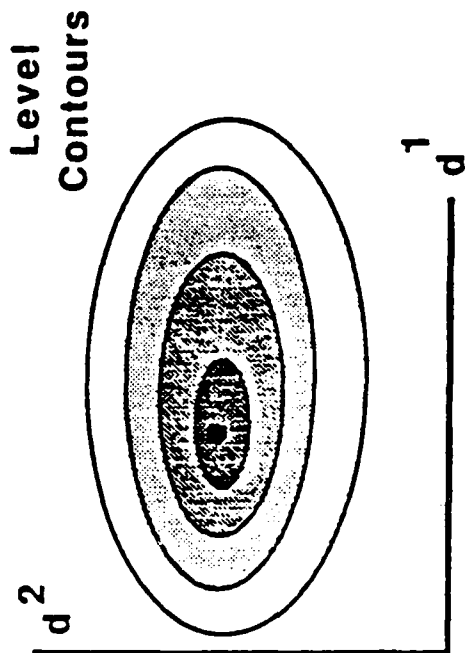
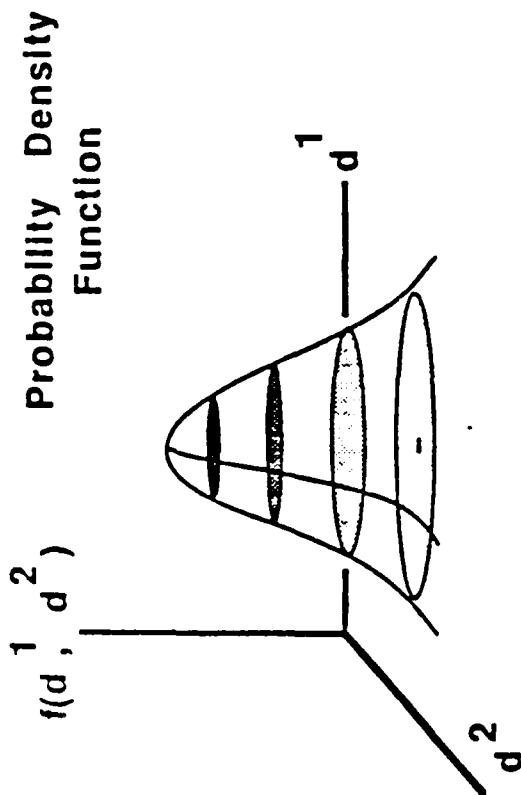
Low Yield Placement



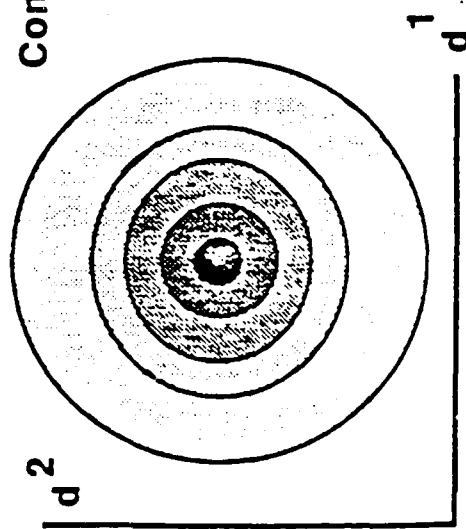
High Yield Placement

$$\text{Yield} = \int_A f(D) dD$$

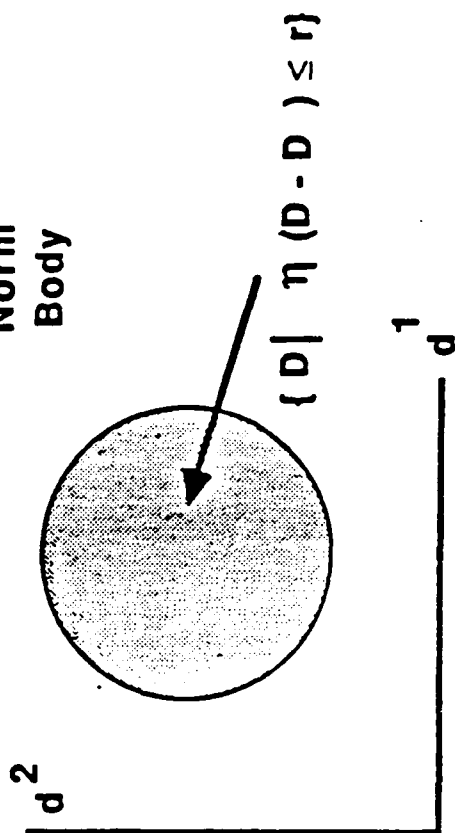
PDF's, Level Contours and Norm Bodies



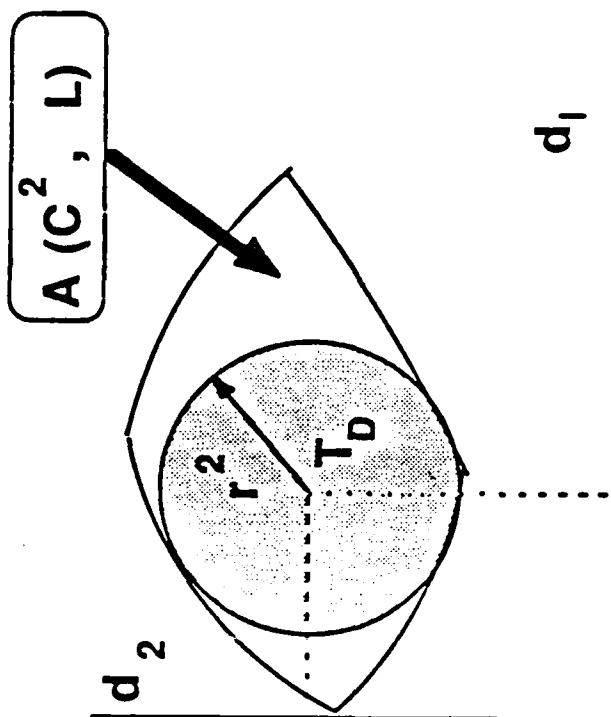
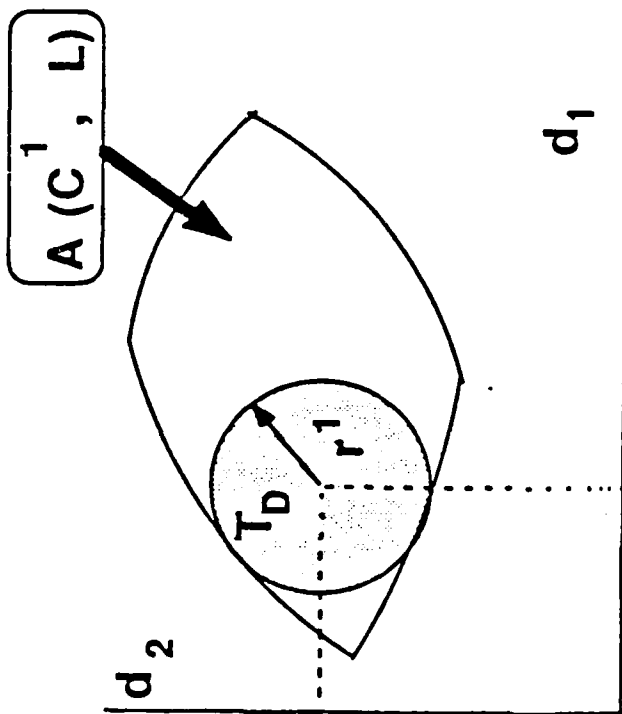
Normalized Level Contours



Norm Body



Process Yield Maximization Problem

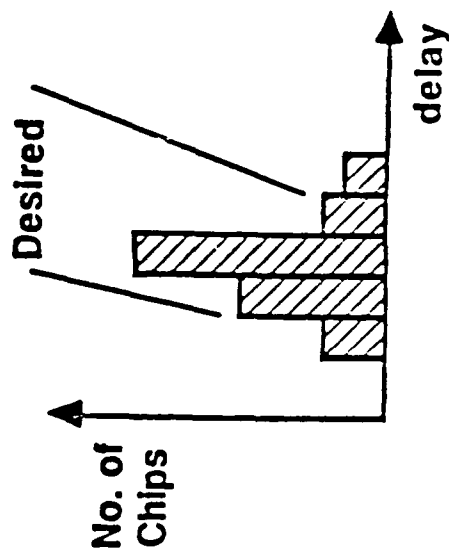
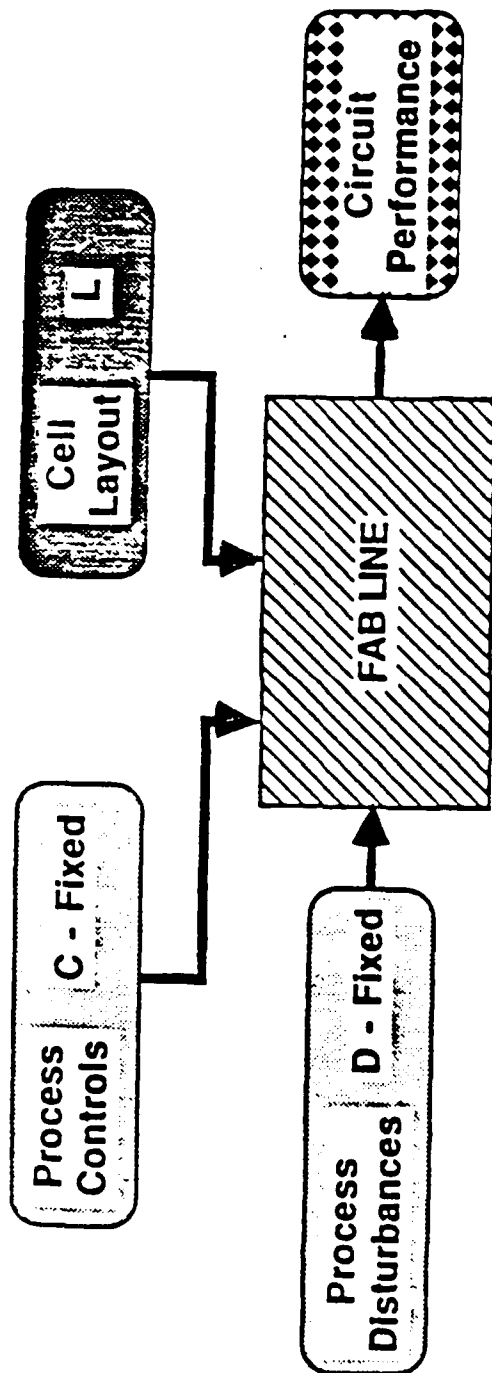


$\max_{C,r} r$

subject to: $T_D(r) \subseteq A(C)$

where $T_D(r) = \{d \mid \eta(d - d^0) \leq r\}$

Cell Optimization



Multilevel Optimization

Designable Parameters:

Local - layout of individual cells

Global - fabrication process controls
common to all cells

Yield Maximization Problem:

Given the JPDF of Process Disturbances

max YIELD

P, L

subject to box constraints on:

P (process control capabilities)

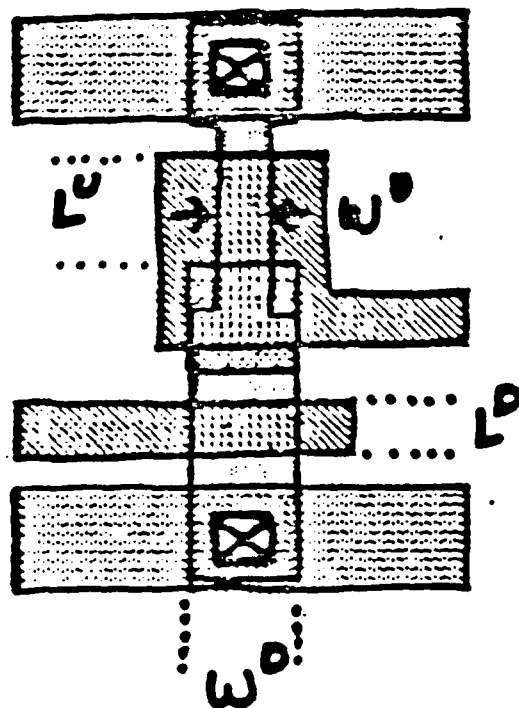
L (lithography capabilities & chip area)

Example

Building macromodels for minicell, and using models to perform *local* optimization of minicell performances.

Problem

Minimize Power-Delay product of an NMOS inverter.



Designable Parameters

- Gate Oxidation Time, T_{ox}
- Depletion Threshold Dose, Q_D
- Minimum Dimension, m
- Ratio of Pullup/Pulldown, R

Performances

1. Power Dissipation, P
2. Rise time, τ_R , (10% to 90% of v_{dd})
3. Fall time, τ_F

Test Conditions

Loaded with 0.01 pF, ~ 1 gate load.

Optimization

Problem: min $\mu(P)\mu(\tau_r)$

X

Constraints:

box constraints on designable parameters:

$$0.75 \leq T_{ox} \leq 1.5$$

$$0.75 \leq Q_D \leq 1.5$$

$$0.75 \leq m$$

$$0.75 \leq R \leq 2.0$$

constraints on performances:

minicell area: $m^2(4 + R) \leq 24$

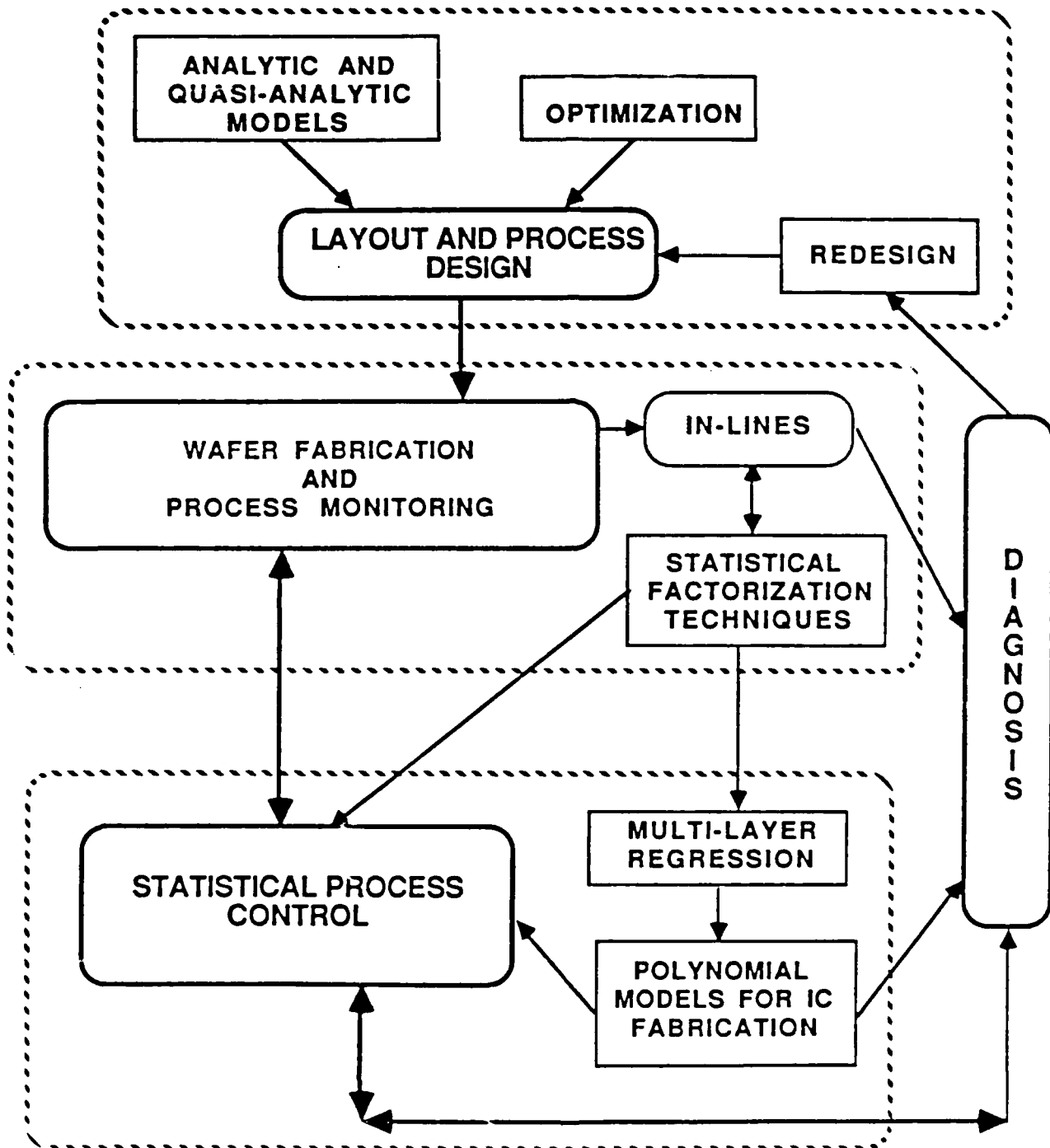
power: $\mu(P) + \sigma(P) \leq 1$

delay: $\sigma(\tau_r)/\mu(\tau_r) \leq 0.1$

Solution:

$$T_{ox} = 1.17 \quad Q_D = 0.96 \quad m = 0.75 \quad R = 0.873$$

CMU-CAM SYSTEM



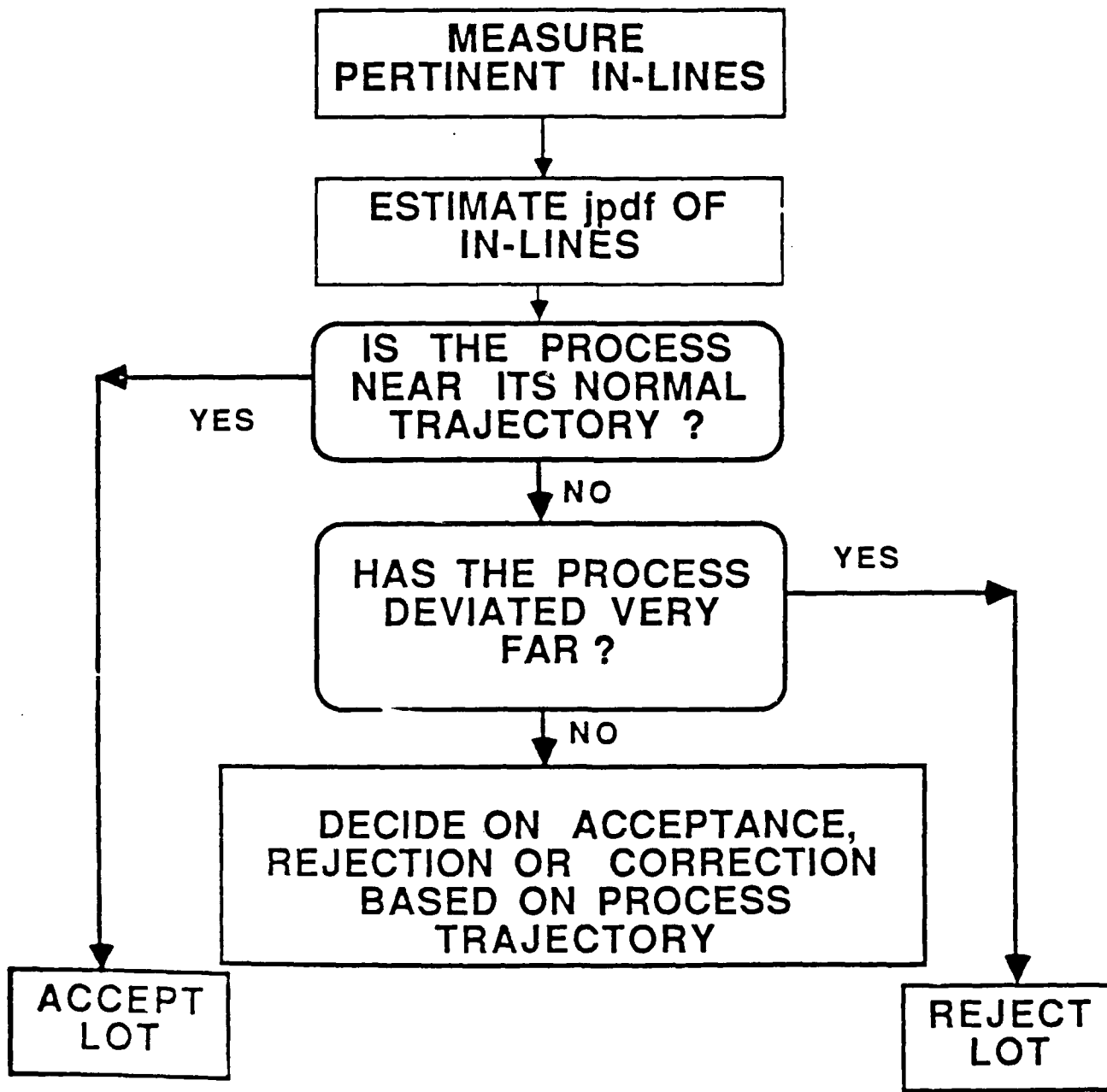
PROCESS OBSERVABILITY

- Measurement types
 - in-line measurements
 - test structure measurements
 - probe measurements
- In-line tests (CD, layer thickness, sheet resistance)
- Scribe lane test structures (performance evaluation)
- Probe measurements (until first fail)
- Needs for process control:
 - increase number of in-line measurements
 - establish relationships between in-line distributions and yield
 - determine selection thresholds and quality control procedures

STAT. QUALITY CONTROL

1. Continue processing if predicted yield $>$ *Threshold of acceptability*
2. Corrective measures \rightarrow *Feed forward control*
3. *Rework*
4. Reject the lot if predicted yield $<$ *Threshold of rejection*
(i.e. further processing is not cost effective)

STATISTICAL QUALITY CONTROL FLOW



PROBLEM DECOMPOSITION

- Shared information between observed in-lines
→correlation
- Grouping in-lines so that no shared information
within groups →clustering
- In-lines in a cluster depend on restricted set of
process parameters and disturbances
- Minimal set of in-lines within each cluster that
needs to be observed →factorization →principal
components
- Identification of *factor in-lines*

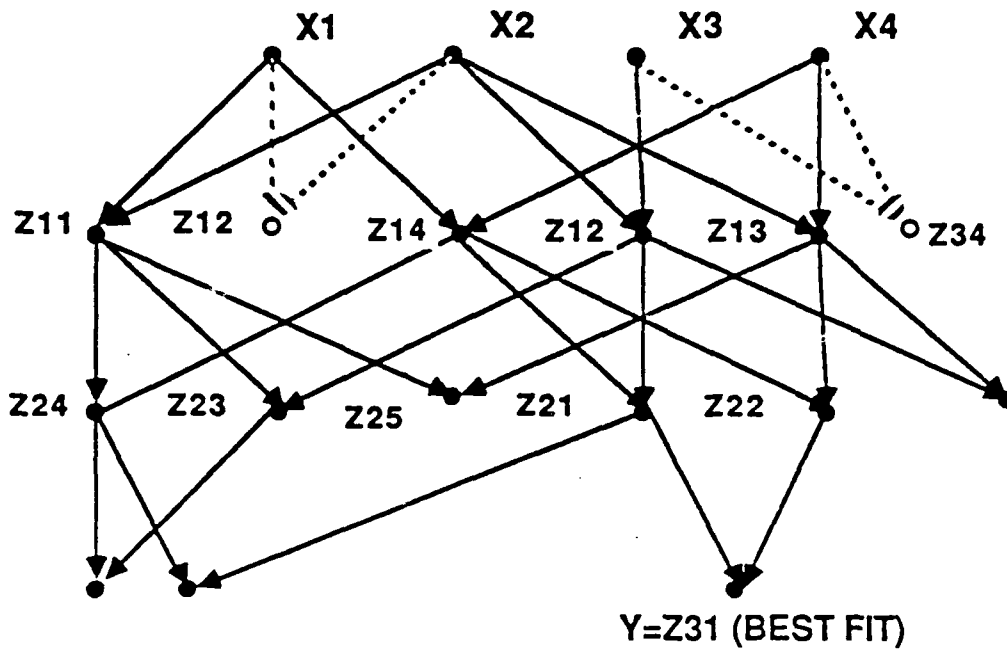
REGRESSION

4 INPUTS

LAYER 1
4 VARS.

LAYER 2
6 VARS.

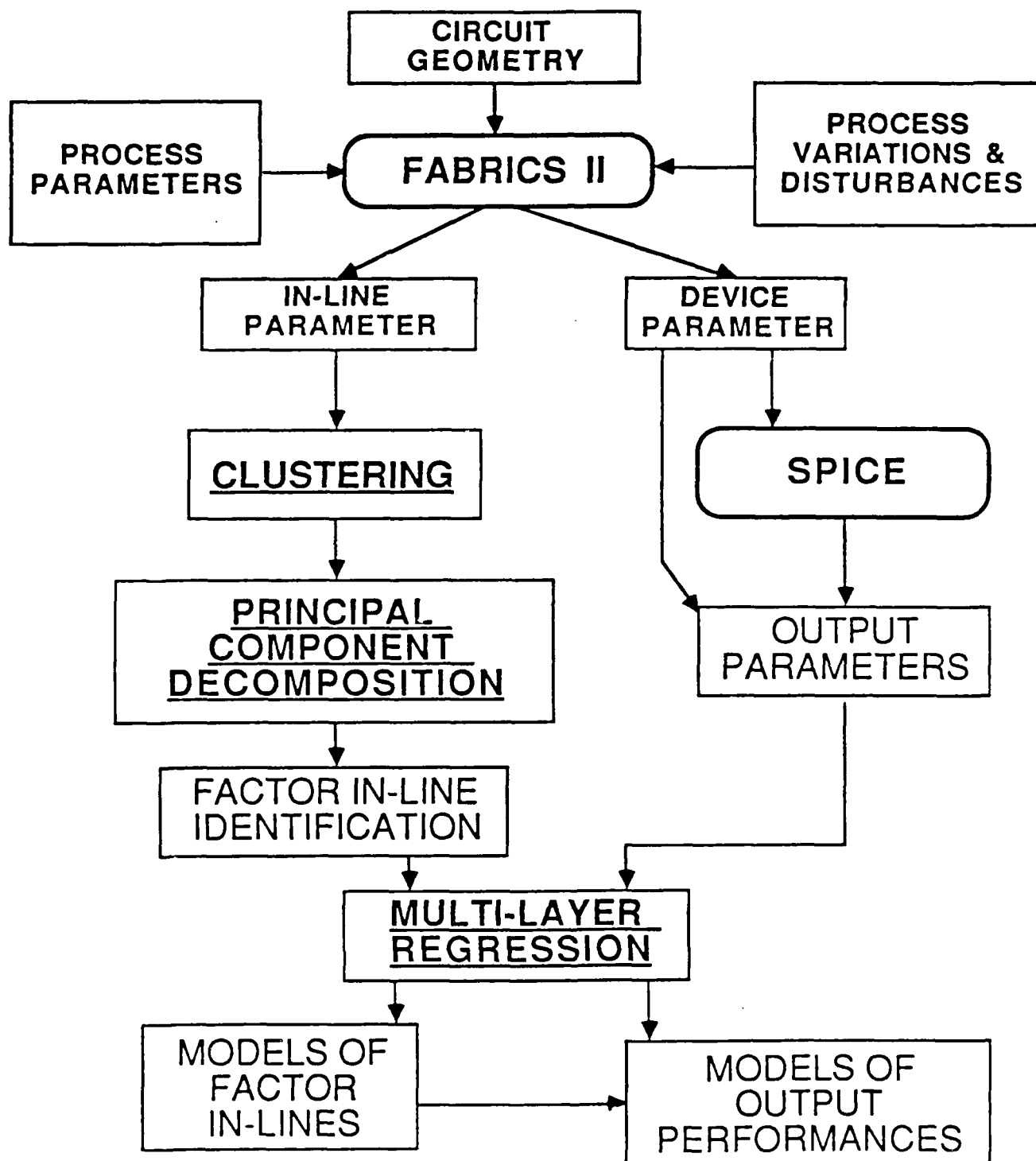
LAYER 3
FINAL LAYER
1 OUT OF 15
CHOSEN



- INTERMEDIATE VARIABLE REJECTED
- INTERMEDIATE VARIABLE CHOSEN FOR NEXT LAYER

- Level 1: P, D → Factor in-lines
- Level 2: Factor in-lines, P, D → X and Non-factor in-lines

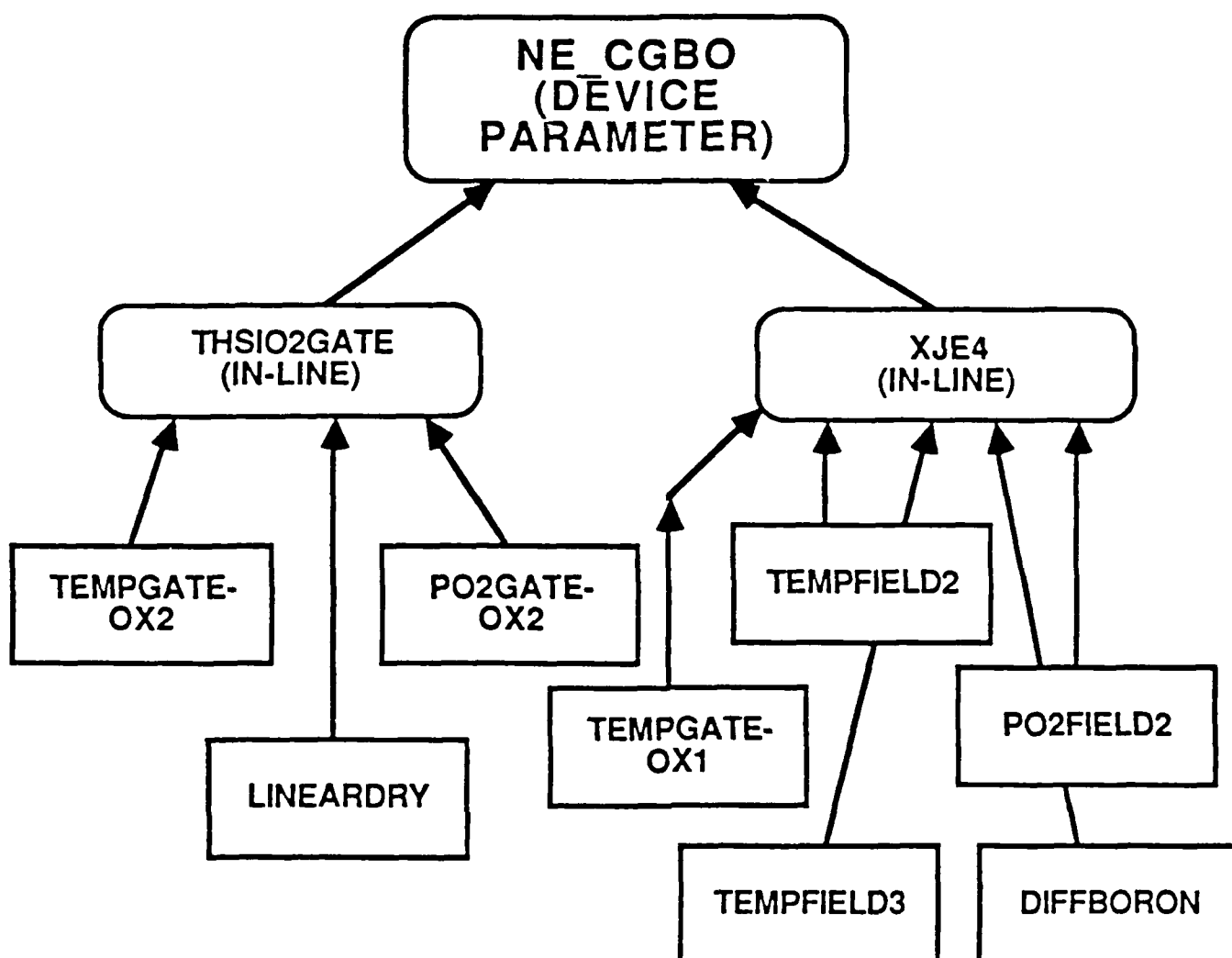
PROCESS MODELING

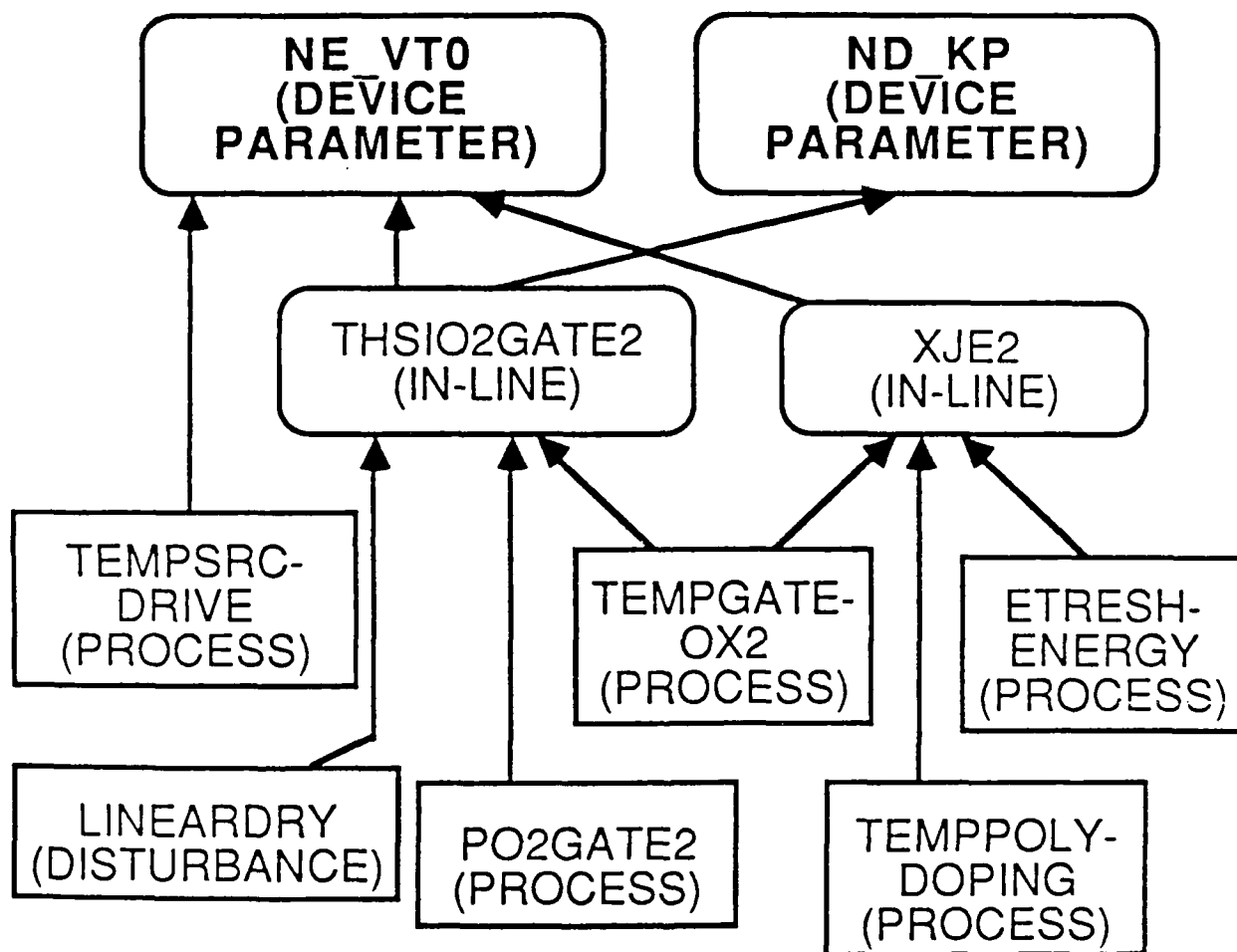


EXAMPLE DECOMPOSITION

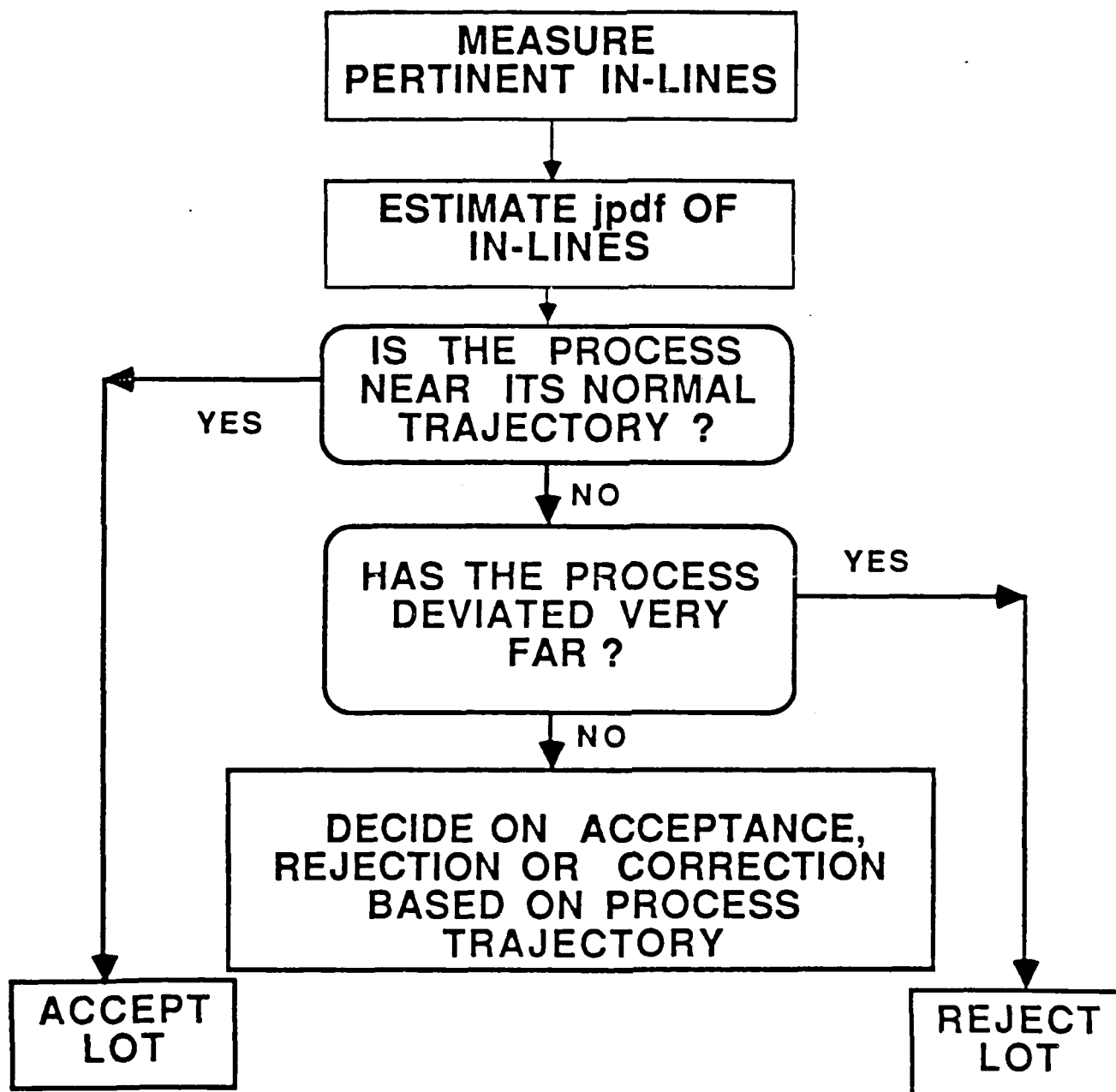
- 4 device NMOS process
- Data generated by FABRICS
- Clustering with threshold of 0.1: 10 distinct clusters
- Principal component decomposition: a few factor in-lines
- 3 CMOS processes from TI →tuned to FABRICS →to be used for further study

2-LAYERED DEVICE MODELS





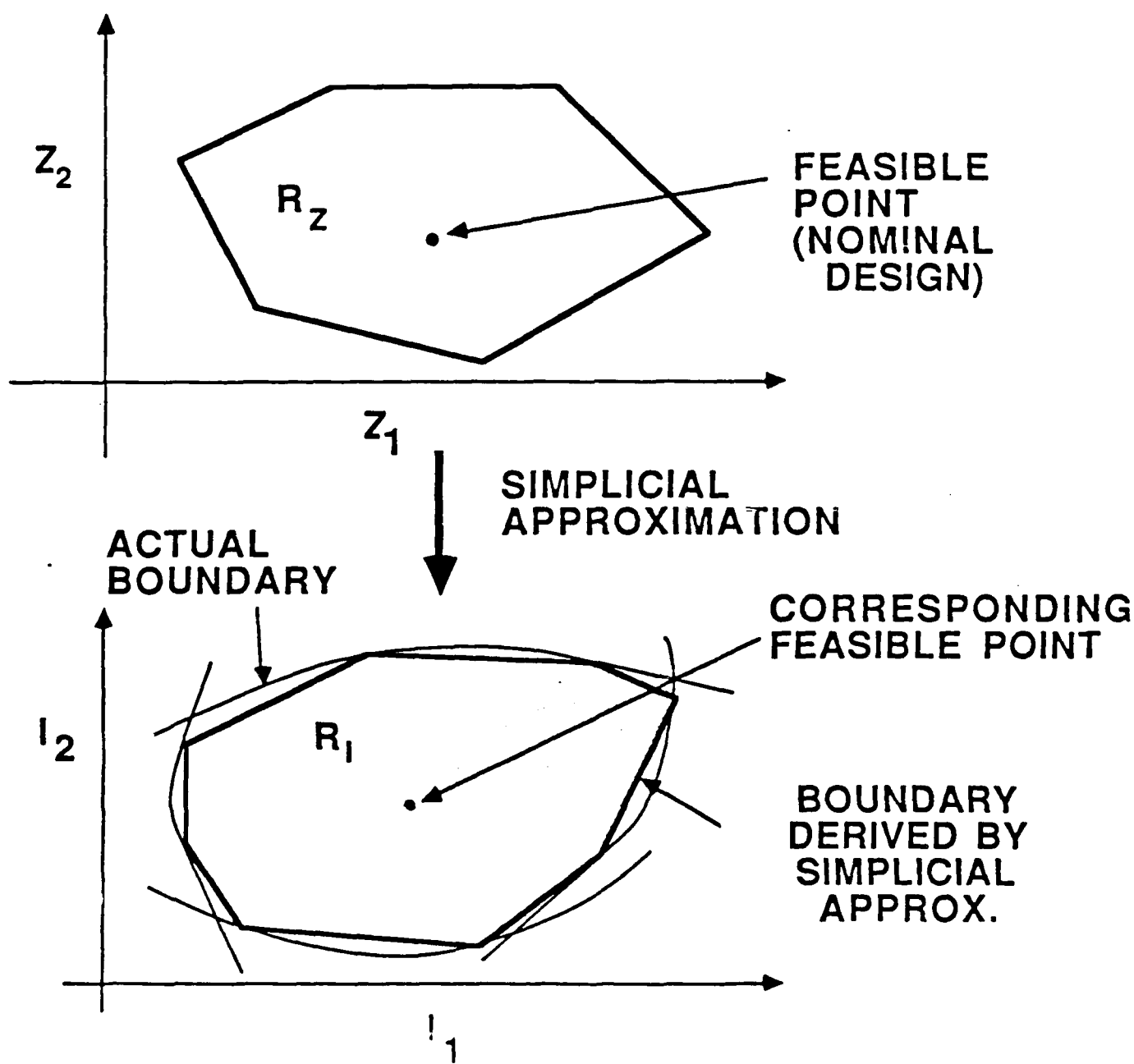
QUALITY CONTROL FLOW



SIMPLICIAL APPROXIMATION

- Acceptability region specified in terms of circuit performances → constraints
- Joint distribution in terms of in-lines
→ estimated during process
- Map back acceptability region to in-lines
→ Simplicial approximation

SIMPLICIAL APPROXIMATION

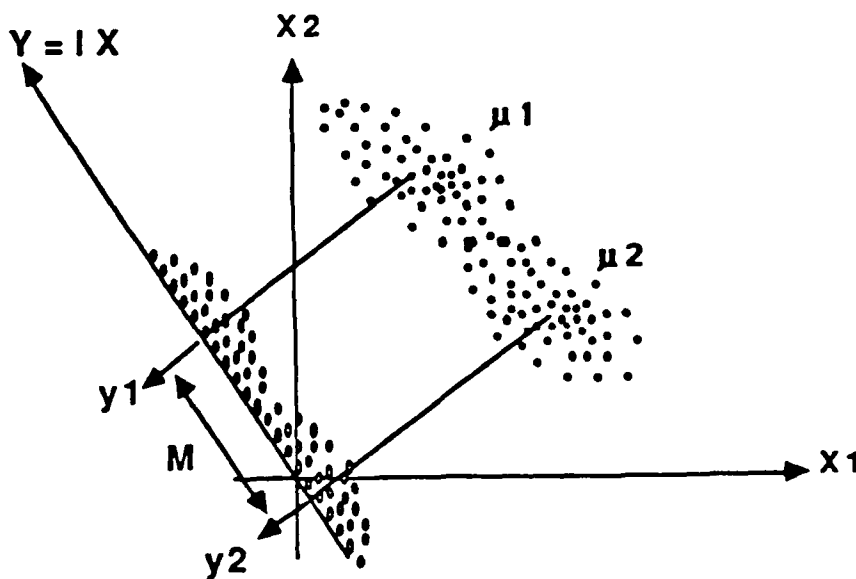


QUALITY CONTROL DECISIONS

- Determine tolerances on distributions → use of process trajectory
 - Acceptance
Statistical distances to determine how far process has deviated
 - Rejection
Moments of distributions of single in-lines
Partial yields

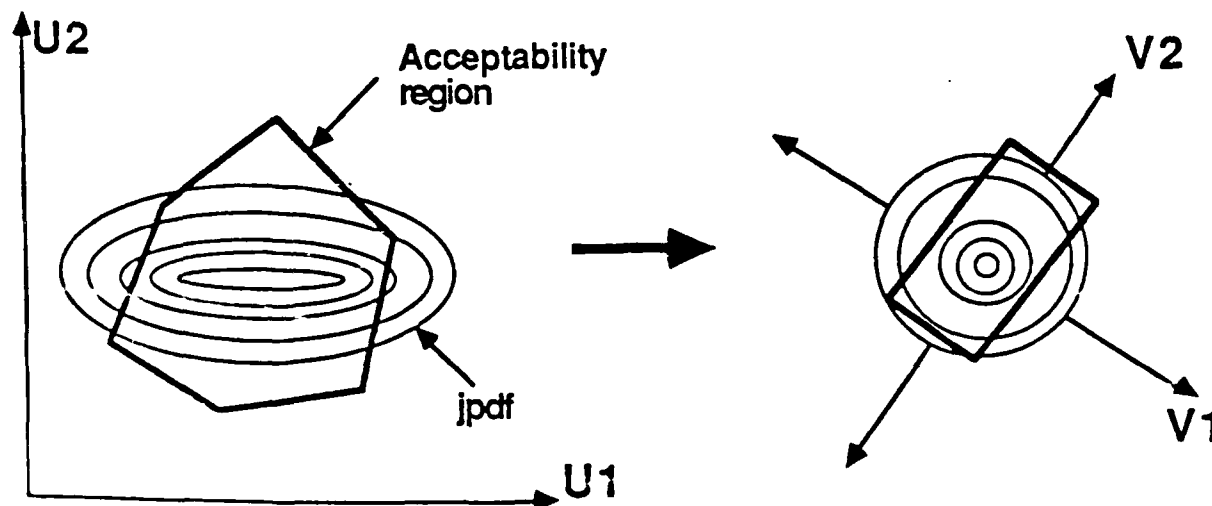
QUALITY CONTROL contd. ACCEPTANCE CRITERION

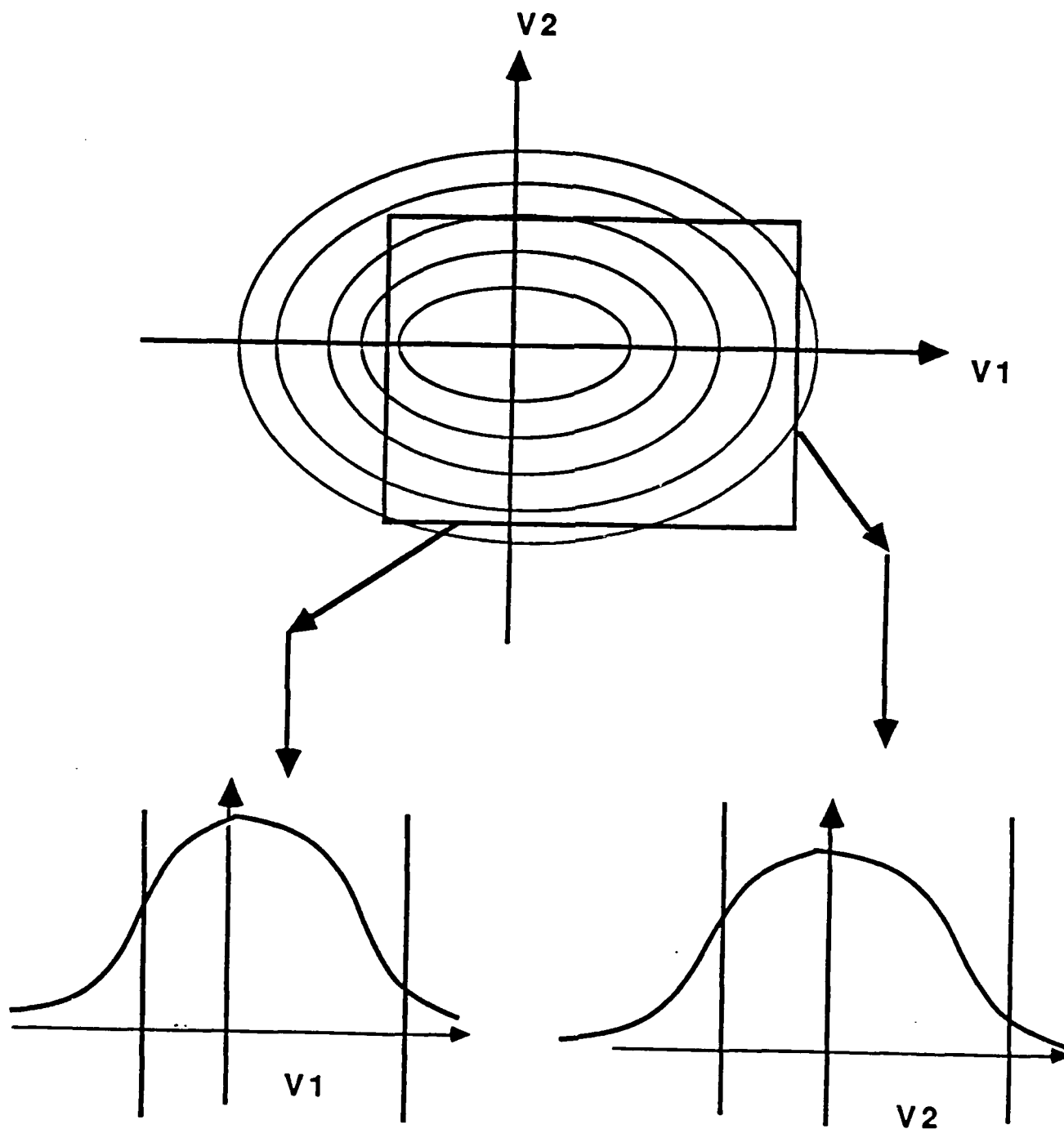
- Distance between jpdfs using nonparametric techniques
- Yield sensitivity to magnitude and direction of shift
- Change in yield due to the shifted mean (e.g. weighted Mahalanobis distance)
- Accept if change is small



QUALITY CONTROL contd. REJECTION CRITERION

- Dimension reduction by factorization (quasi-independence)
- Simplified approximation of acceptability region by hyperbox
- Appropriate coordinate transforms
- Rejection based on single in-line distribution
→ partial yields → tolerances on in-line distribution



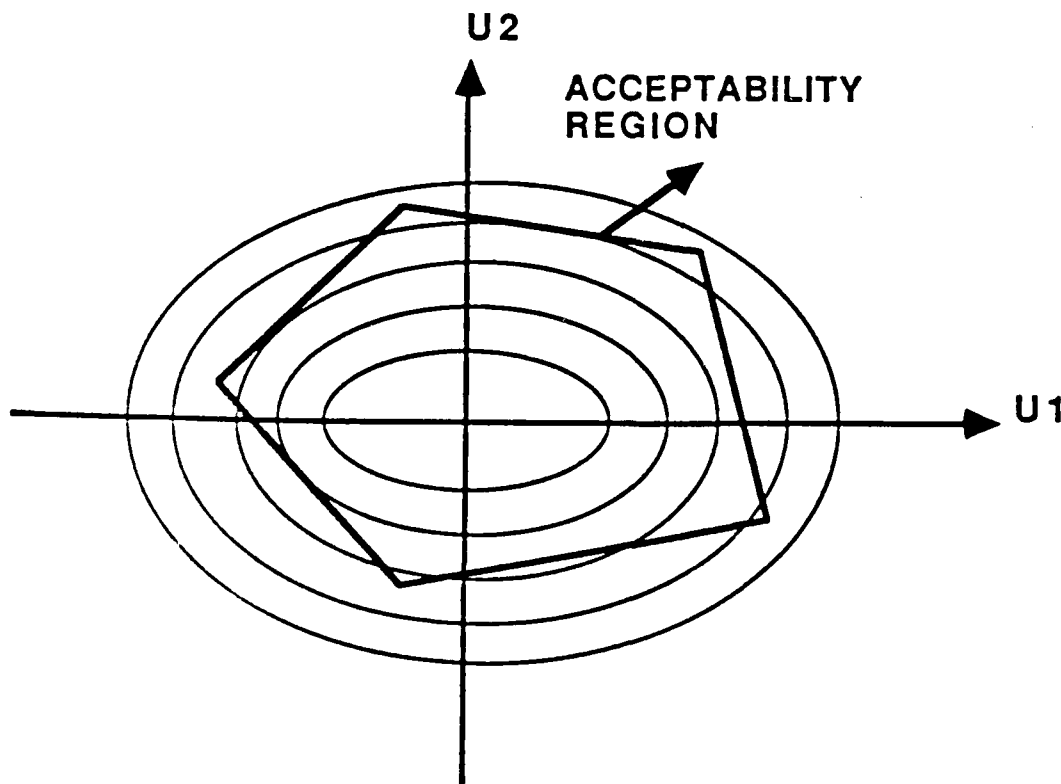


YIELD PREDICTION

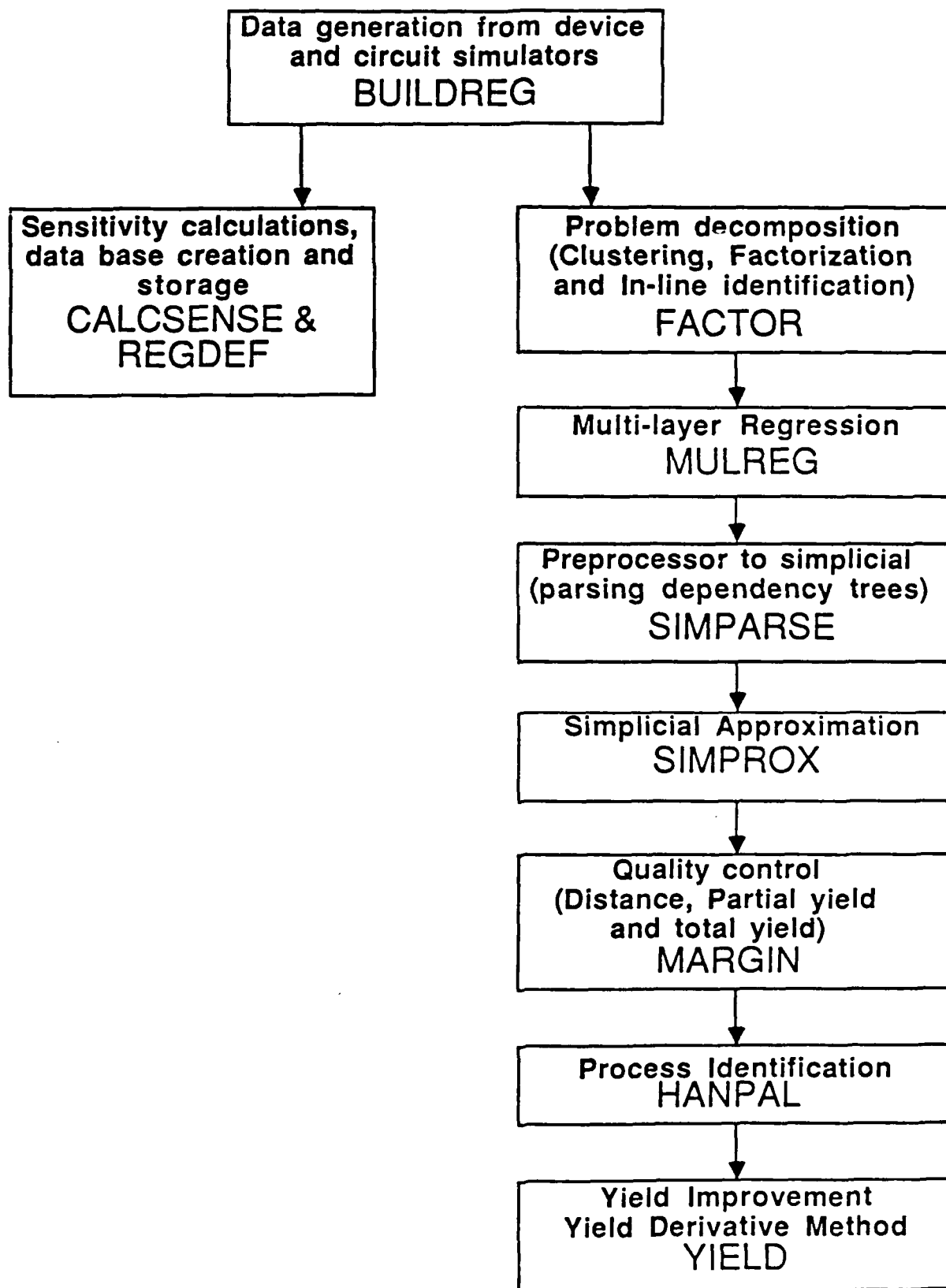
- Integrating jpdf over acceptability region in in-line or circuit performance space

$$Y = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \phi(U) \psi_U(U) d_{u_1} \dots d_{u_{nu}}$$

- Coupling → large computing time for integration
- Low dimensionality for control



SYSTEM OVERVIEW



EXAMPLE

- Parameter: NE—CGB0, Acceptance level: 85%
- Observables: thsio2gate, xjf4
- Acceptance:
 - Partial yield due to xjf4 = 99%
 - Partial yield due to thsio2gate2 = 93.8%
 - The Mahalanobis distance between the two distributions = -6.3005e-03
 - The estimated yield is = 93.0%
- Rejection:
 - Partial yield due to xjf4 = 99%
 - Partial yield due to thsio2gate2 = 77%
 - The Mahalanobis distance between the two distributions = 9.7272e-02
 - The estimated yield is = 79.5%

EXAMPLE

- Parameters: NE—VT0 and ND—KP,
Acceptance level: 75%
- Observables: thsio2gate, xje2, tempsrcdrive
- Partial yield due to thsio2gate2 = 79%
Partial yield due to xje2 = 98%
The Mahalanobis distance between the two
distributions = $2.66e-2$
The estimated yield is = 71%

CURRENT WORK

Statistical simulation

- sampling techniques**
- analytical mapping of jpdf's**

Yield prediction

- dimensionality reduction techniques**

Statistical design

- inverse mapping techniques**
- factor-splitting approach**

Statistical process control

- nonparametric methods**

Statistical Optimization for Computational Models

Kishore Singhal

AT&T Bell Laboratories
1247 South Cedar Crest Blvd.
Allentown, PA 18103

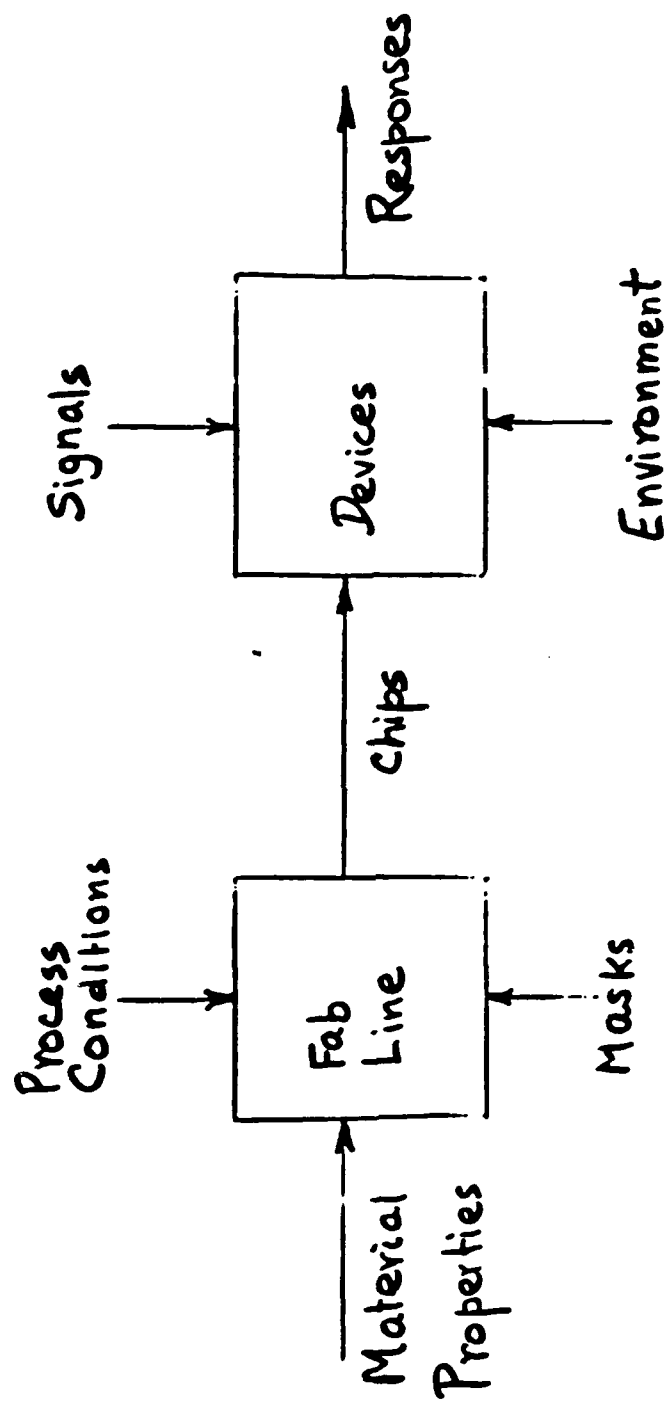
Realistic engineering problems of expected product cost minimization and reliability maximization in the presence of manufacturing fluctuations and parameter variations due to environmental and age related effects can be formulated mathematically as constrained statistical optimization problems. *Parametric Sampling* is a particular technique for solving such problems when the system performance can be obtained through simulation using computational models.

A database containing the results from a small number of simulations is first created. Parametric Sampling allows us to estimate the objective function, the constraints and their gradients not only at the initial set of design parameters but also at new design points generated by the optimization algorithm. As needed, additional sample points are added to the database to ensure estimation accuracy.

Sensitivity studies to determine the influence of specification changes and departures from assumed statistical distributions are possible with minimal computational cost. Experience with electrical and mechanical systems shows that substantial improvement in the objective functions is possible.

Conference on Uncertainty in
Engineering Design, 1988 (May 10, 11)

Ref: Statistical Design Centering and Tolerancing Using Parametric Sampling. IEEE Trans. Circuits and Systems, July 1981.

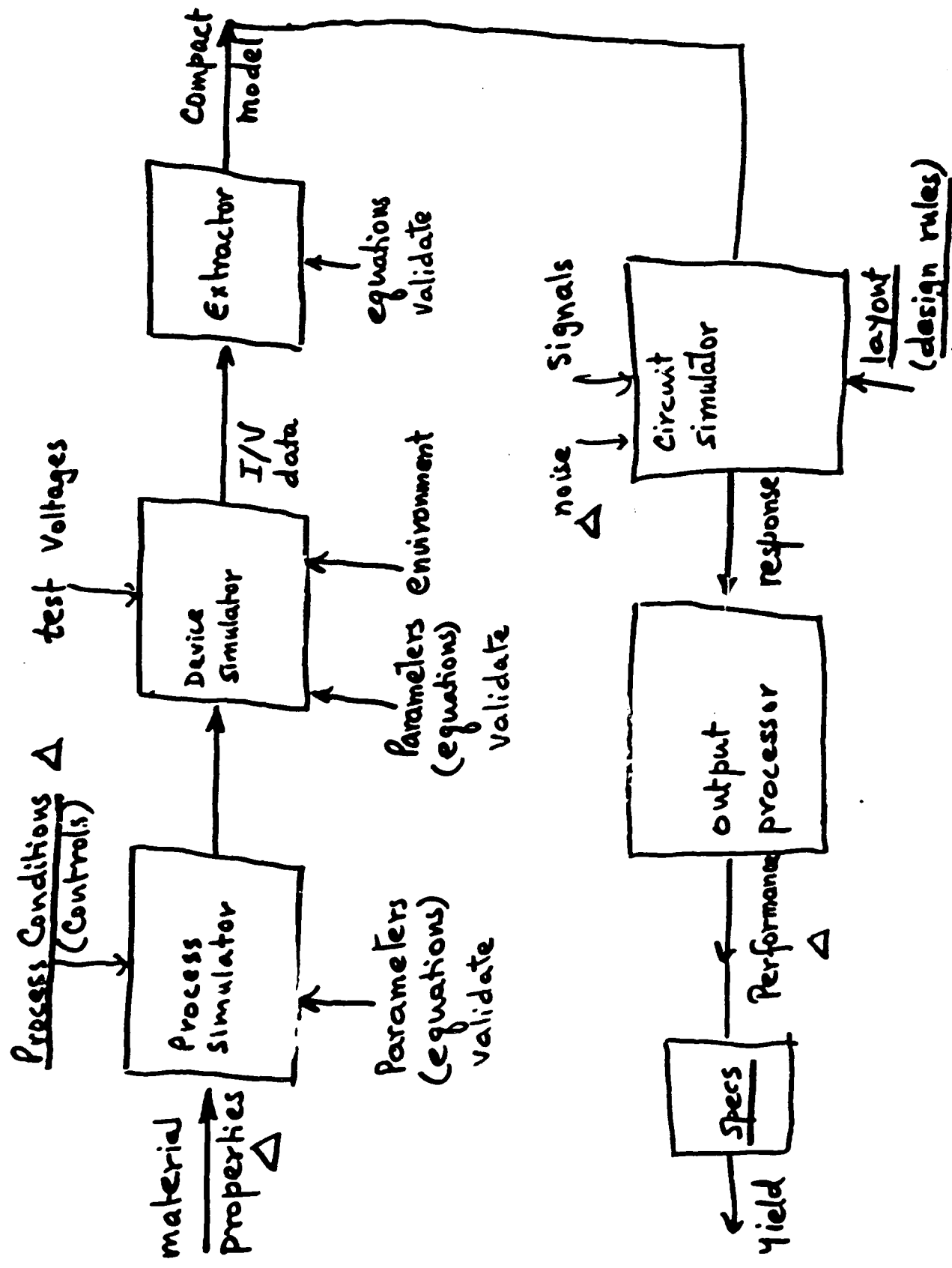


Computer Simulation vs Physical Experimentation

- + Often faster (real time)
- + Less expensive
- + Easy to explore areas of operation where physical experimentation may be difficult
 - Need a computer model
 - Need reasonable characterization of probability model

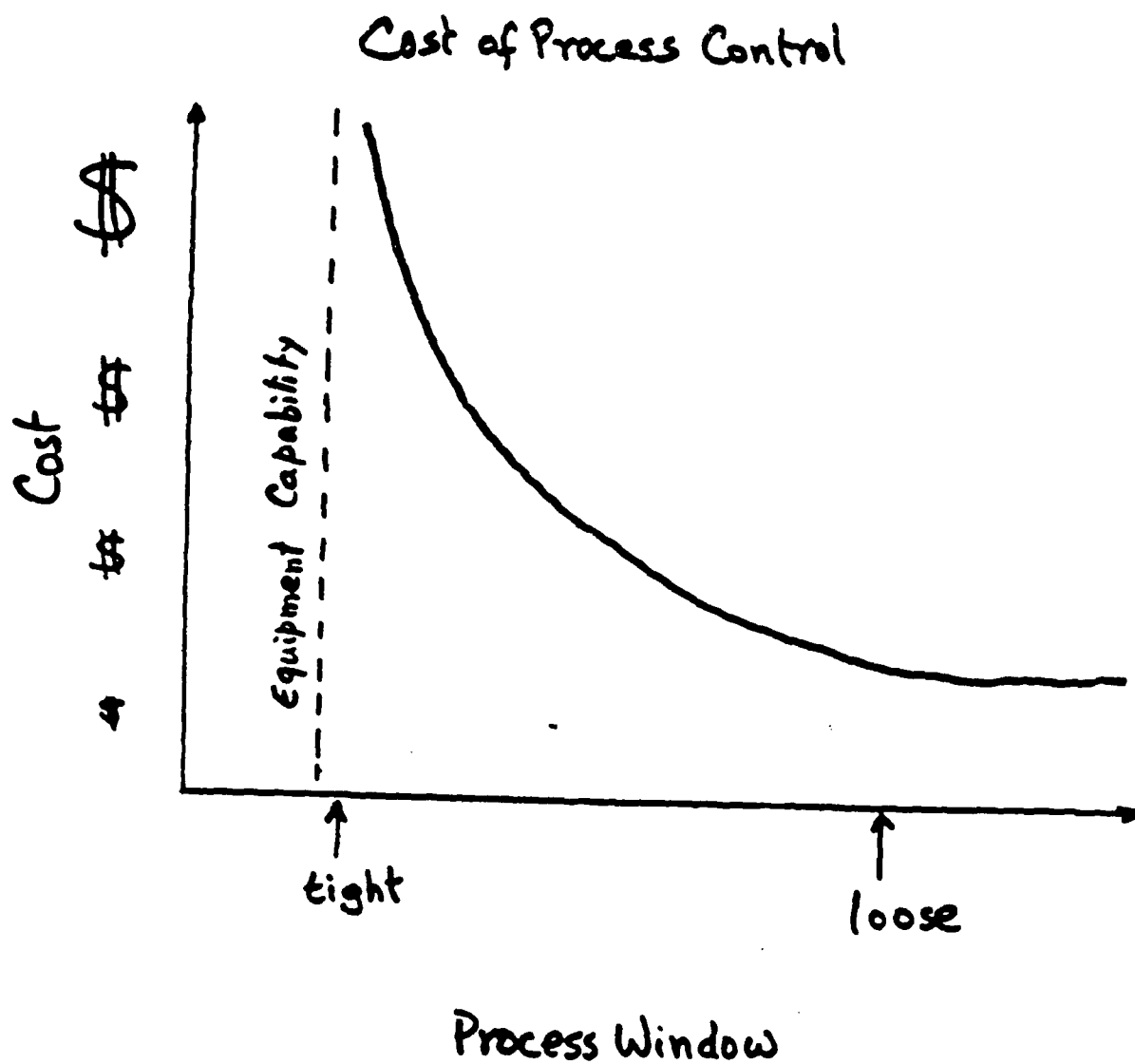
Key difference in Optimization Strategy

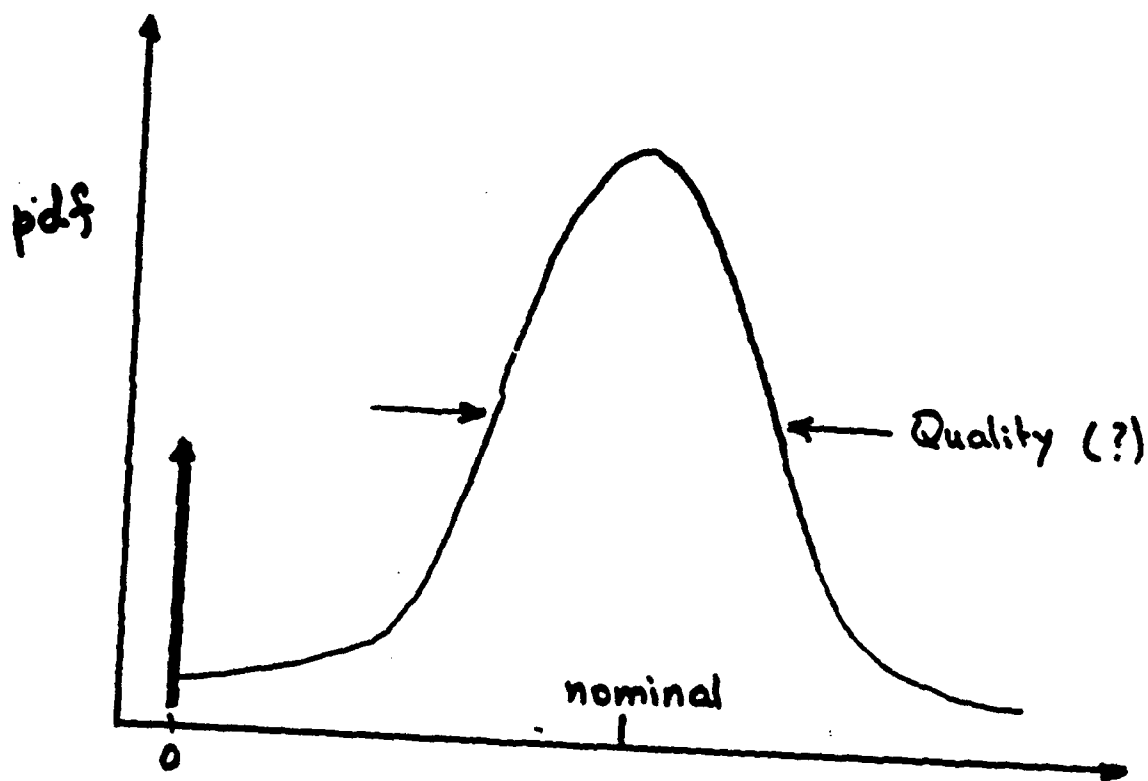
- We can distort reality to assist us
- Analysis and Decision steps can be easily interlaced
- Computational complexity is no longer an issue



- Controls

- Random Variability



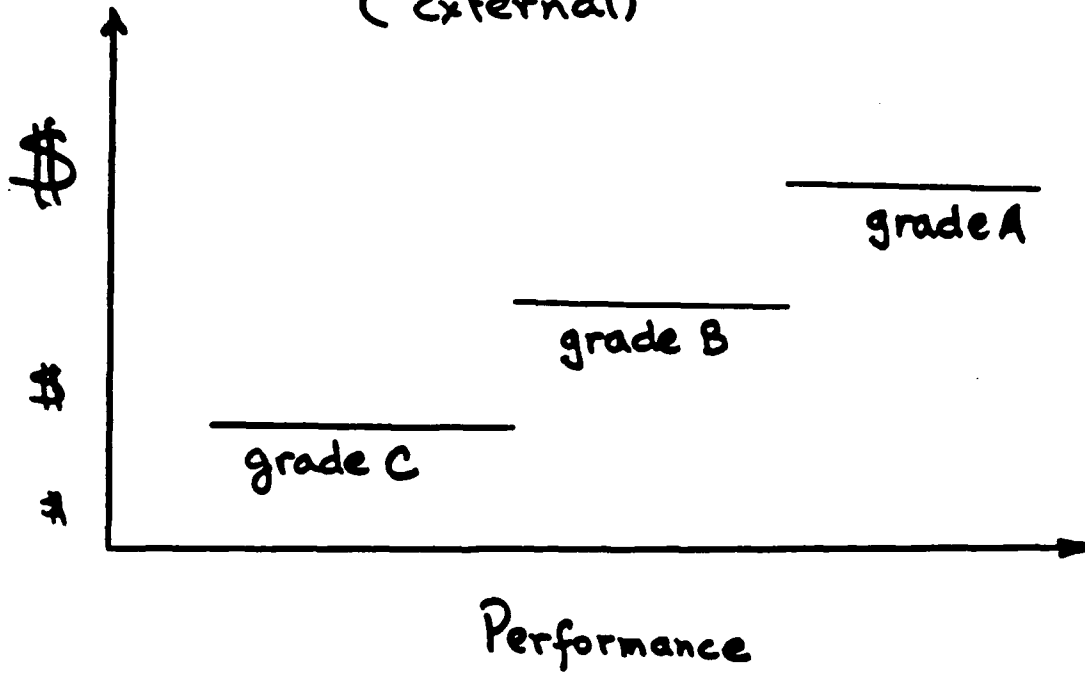


Performance

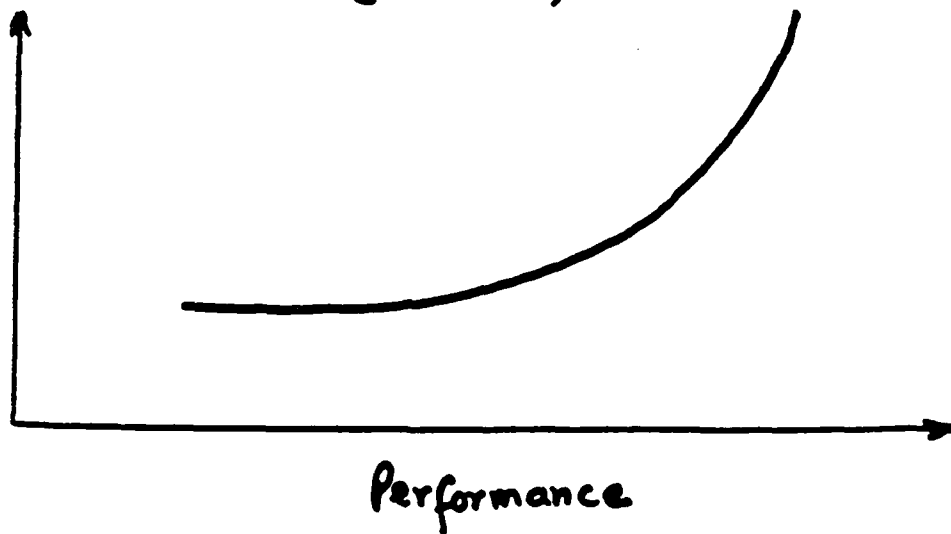
pdf depends on

- Controls
 - Process nominals, windows
 - Circuit design, design rules
- Noise material properties etc.

Sales Price (External)



Value (internal)



Objective

Maximize profits :

Related to

product volume (yield)

Sales price

Cost of Control

Design time

Reliability

⋮

By selecting

Process nominals
windows

Design Rules

} Tech.

Circuit Design.

Mathematical problem translates to
Computing high dimensional integrals
of the form

$$\int v(x) f(x; \theta) dx$$

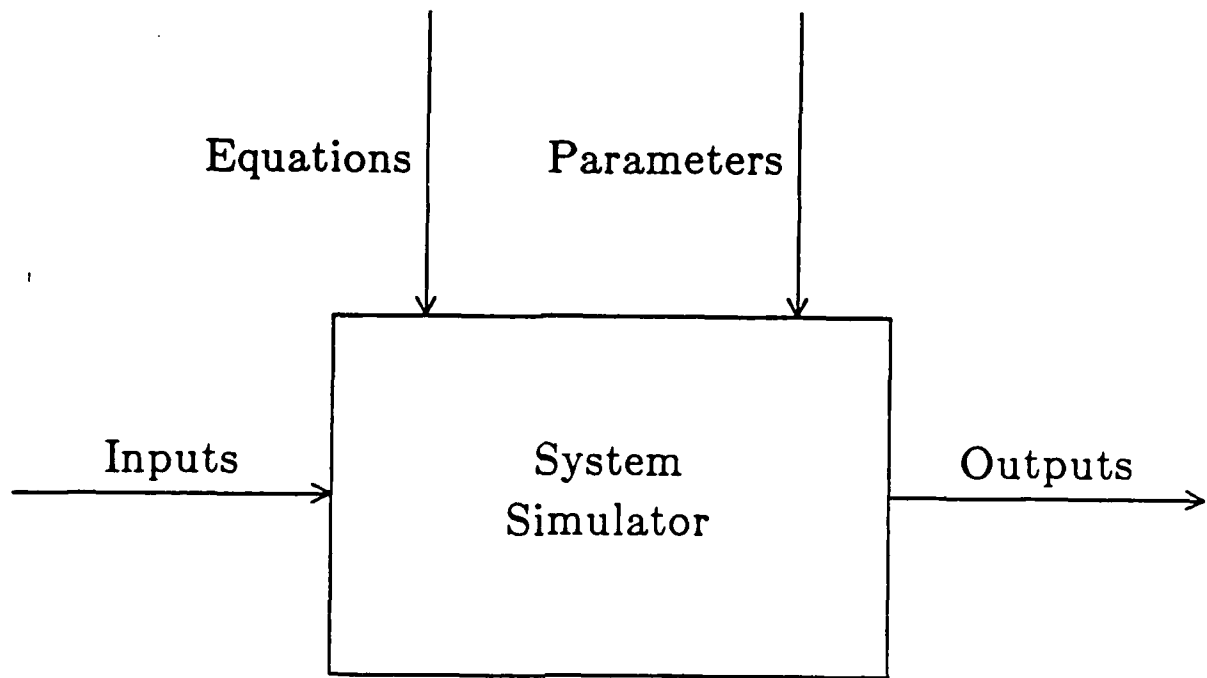
Design for Manufacturability, Reliability and Minimum Product Life Cycle Cost

Include

- Process variation
- Tolerance vs cost tradeoffs
- Environmental condition variations
- Limits on performance degradation with age
- Testing and field repair costs
- Cost of lost goodwill
- etc etc etc

in the design phase itself

Simulator Components



Statistical Simulator Components

Distribution $f(x;\theta)$

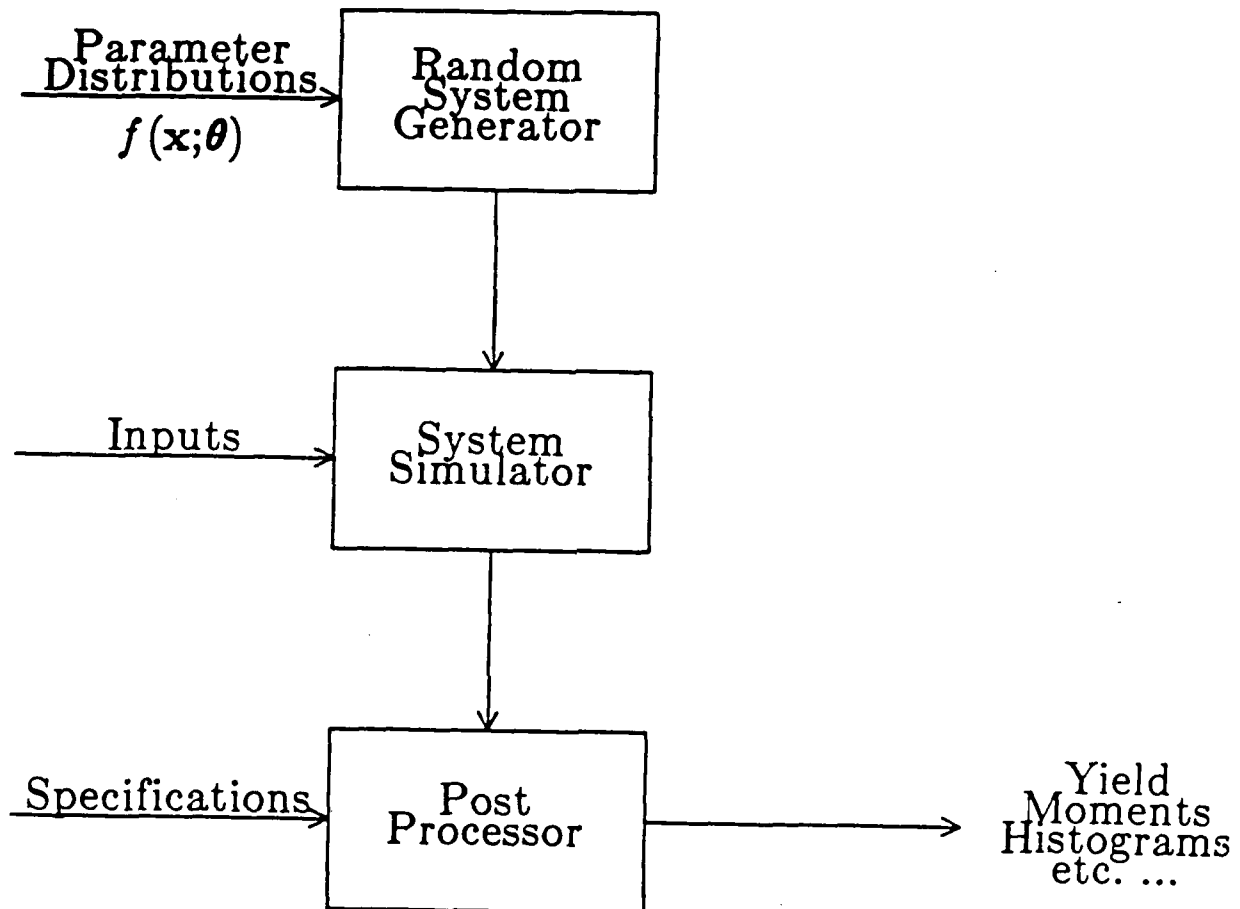
$$\theta = \begin{cases} \textit{nominals} \\ \textit{tolerances} \\ \textit{correlations} \end{cases}$$

Random
variables

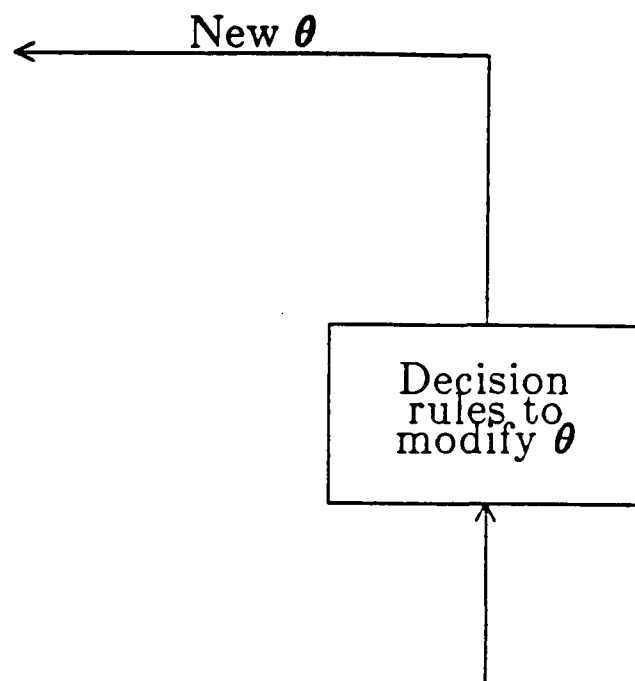
↑
Noise
Temperature
Humidity
Age ..

Simple Monte Carlo

11a

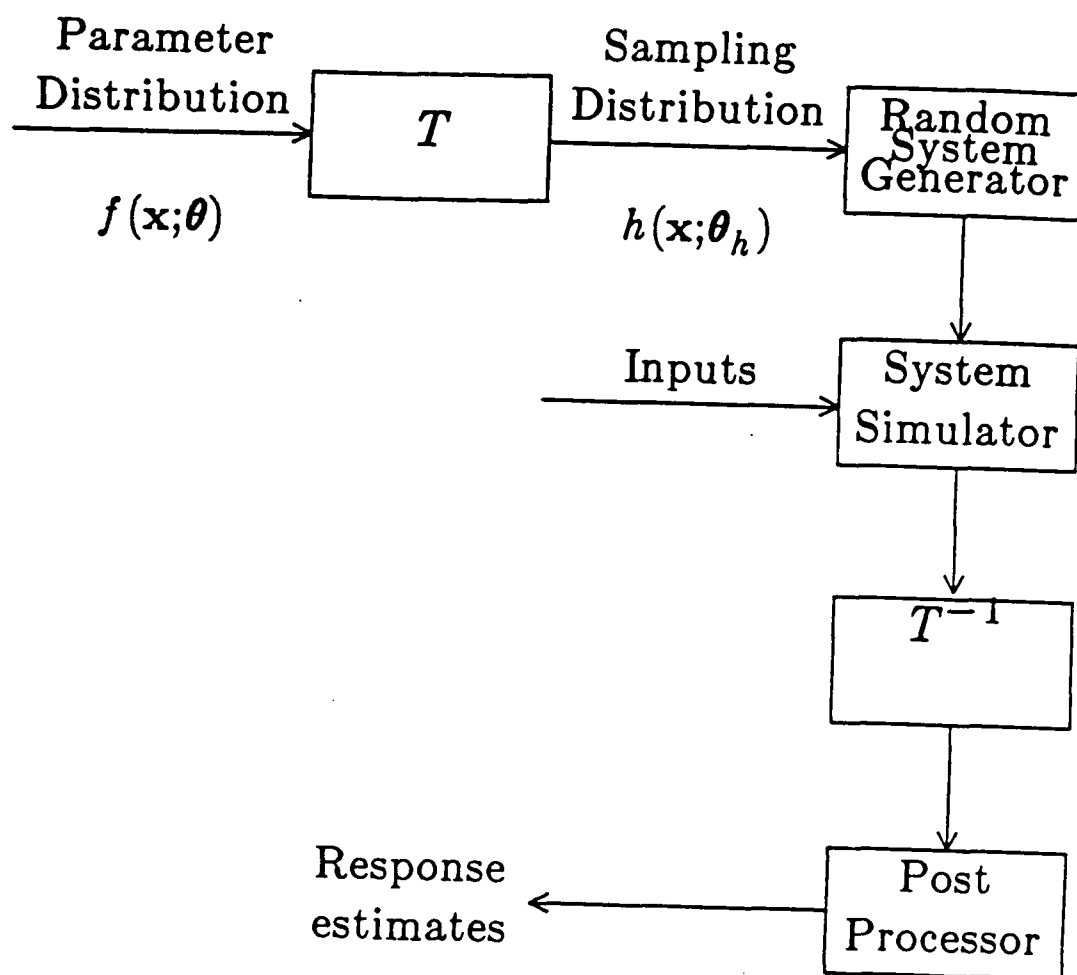


Simple Statistical Design

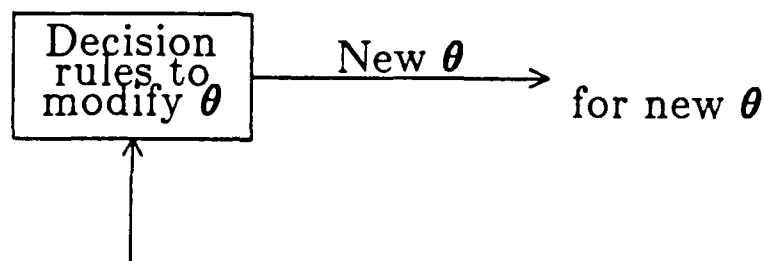


Variance Reduction by Importance Sampling

12c




Statistical Design by Parametric Sampling



Statistical Design by Parametric Sampling

new
 $h(\mathbf{x}; \theta_h)$



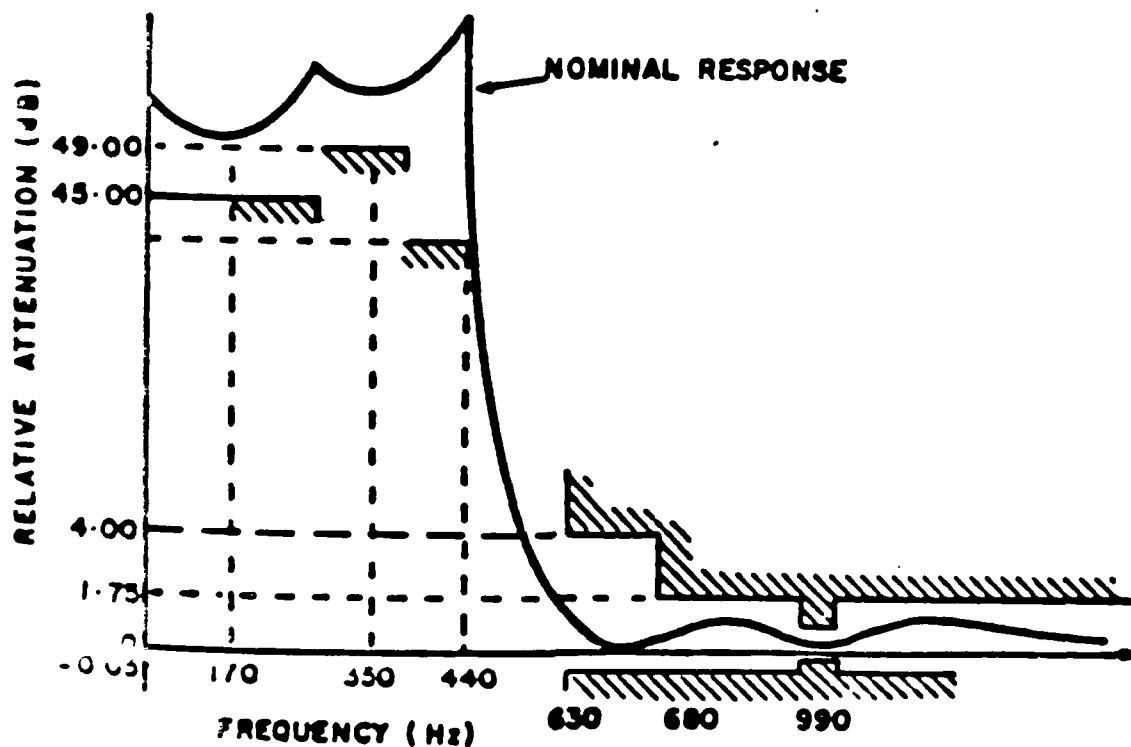
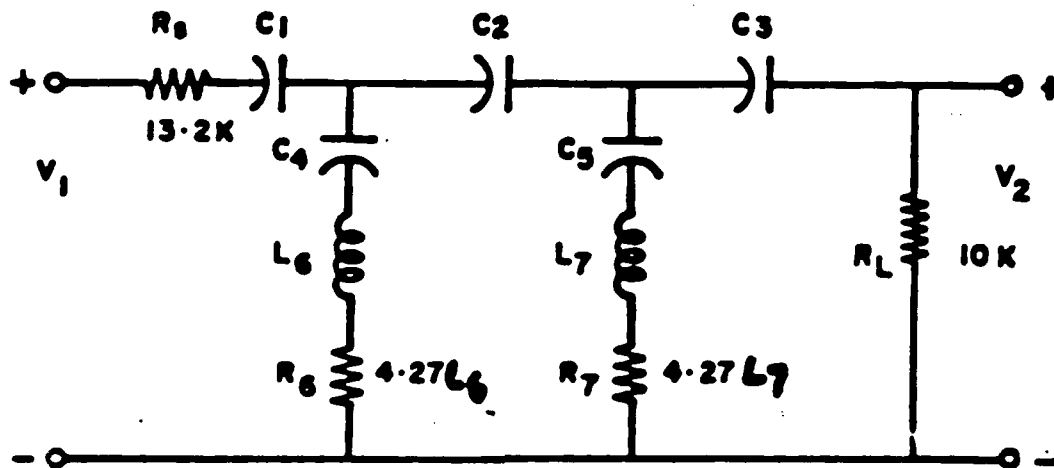
High Pass Filter Problem

13

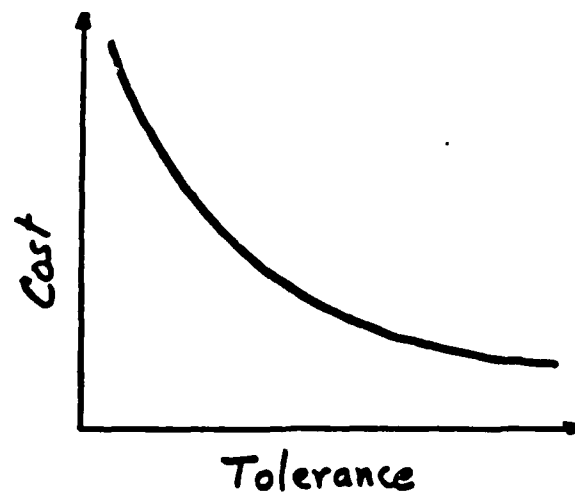
Example

$$C(t) = 2.5 + \sum_{i=1}^5 \frac{1}{t_i} + \sum_{i=6}^7 \frac{2}{t_i}$$

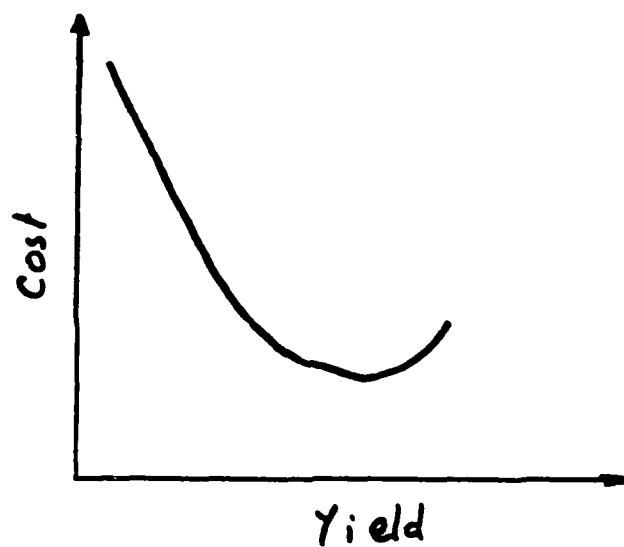
$$C_s(\ell) = \frac{C(t)}{y(\ell)}$$



High-Pass Filter Specifications.



$$C_i \propto \frac{1}{t_i}$$



$$\text{Cost} = \frac{\text{Component cost} + \text{fixed cost}}{\text{Yield}}$$

- starting from nominal design and $t = 10\%$

	initial	statistical design
yield	86.7	94.9
Cost	3.92	3.52

192 ckt analyses

- starting from worst case centered design

	initial	statistical design
yield	100	96.9
Cost	4.53	3.32

160 ckt analyses

Ordinary Monte Carlo

Evaluate high dimensional integrals of the form

$$E(v) = \int_{-\infty}^{\infty} v(x) f(x; \theta) dx$$

where v is a function computed through simulation and $f(x; \theta)$ is a density with parameters θ by sampling as

$$\hat{v} = \frac{1}{N} \sum_{i=1}^N v(x_i)$$

where x_i are samples drawn from the distribution $f(x; \theta)$

Importance Sampling

$$E(v) = \int_{-\infty}^{\infty} v(x) f(x; \theta) dx$$

can be written as

$$E(v) = \int_{-\infty}^{\infty} v(x) \frac{f(x; \theta)}{h(x; \theta_h)} h(x; \theta_h) dx$$

where $h(x; \theta_h)$ is some other density

The second integral can be approximately evaluated by sampling as

$$\hat{v} = \frac{1}{N} \sum_{i=1}^N v(x_i) \frac{f(x_i; \theta)}{h(x_i; \theta_h)}$$

where x_i are samples drawn from the distribution $h(x; \theta_h)$ selected to reduce the variance of \hat{v}

Parametric Sampling

$$\hat{v} = \frac{1}{N} \sum_{i=1}^N v(\mathbf{x}_i) \frac{f(\mathbf{x}_i; \theta)}{h(\mathbf{x}_i; \theta_h)}$$

where \mathbf{x}_i are samples drawn from the distribution $h(\mathbf{x}; \theta_h)$ independent of θ !!

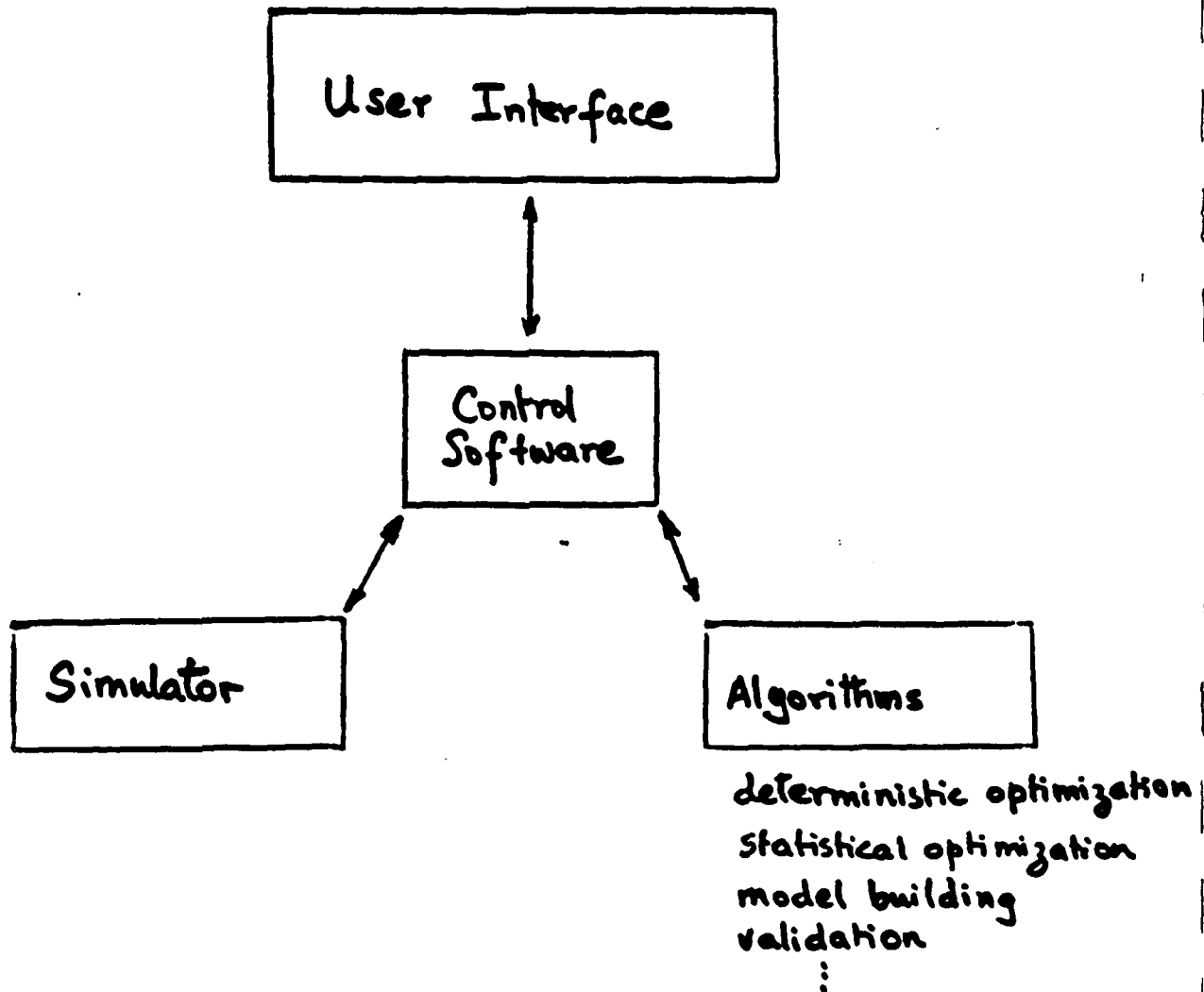
- Functional form of $f(\mathbf{x}_i; \theta)$ is known and \hat{v} can be evaluated for any θ without performing a new Monte Carlo
- First and higher order derivatives of \hat{v} are easily computed allowing the use of powerful optimization tools

Parametric Sampling thus changes a statistical optimization problem into a standard deterministic problem

Some Details

- Stochastic Approximation principles used to force convergence in probability
- The multivariate normal sampling density is related to the Hessian information obtained from the optimizer and enables navigation along narrow ridges
- Quasi random numbers are used to reduce estimator variance
- Sample pooling to reduce sample size
- Ratio estimators improve accuracy
- Jackknife for bias reduction and error estimation
- Sensitivity to parameter distributions easily computed
- Re-design following specification changes is simple

Software Structure.



↔ Data Communication Protocol

- Subroutines
- Files and sub-tasks
- 2-Way pipes and signals
- Shared memory IPC

PROFESSOR JEFF WU
University of Wisconsin

DISCUSSANT NOTES

2
Statistical variations at
every stage of design & manufacturing

- model variations
(statistical, physical modelling,)
simulation
- control variations
accommodate variations
(control charts, sampling inspection,
experimental design, parameter design)
- assess impact
(economic loss, reliability prediction)

In engineering curriculum ?

Some interesting statistical issues
in the presentations:

3

- parameter (importance) sampling
- design centering & tolerancing
- complex model building
(macromodel, hierarchical)
- physical vs computer exp't
⇒ difference in design strategies
- sequential generation of data
- interlaced with parameter optimization

Why model the moments?

+ easier for optimization

- information loss

- nonlinearity & other structures
masked

Why not model the data?

Parameter Sampling :

5



- Choice of θ_{h+1} ?

- Efficient way of pooling data ?

given $(\mathcal{Z}_h | \theta_h)$,

$(\mathcal{Z}_{h+1} = ? | \theta_{h+1})$

} efficient

combination :

Centering & Tolerancing:
multi dimensional



economic way to identify
robust settings

θ_{htl} : robust design (Taguchi)
Traditional experimental design:

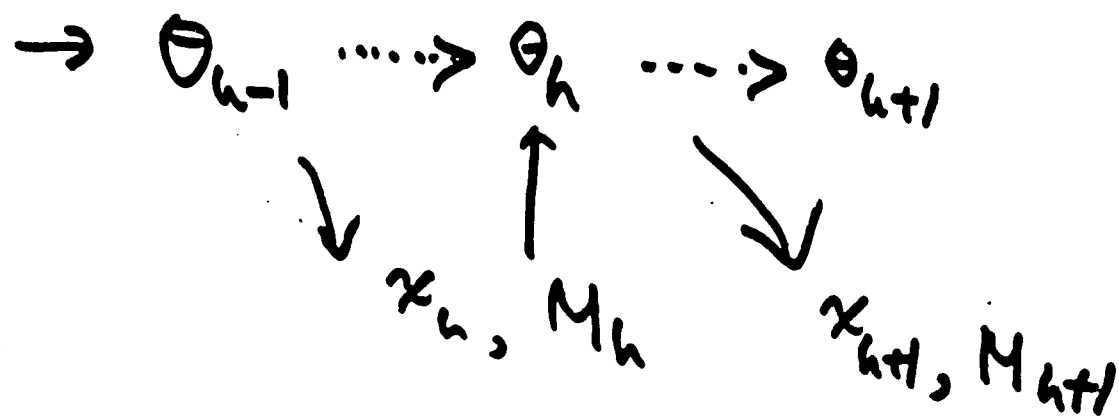
- mean (main effects, interactions)
- variation used as a secondary measure to assess significance of mean

7

Data from investigation of complex system :

- high dimensional
- often expensive \Rightarrow sparse
- availability of reliable
model ??

Sequential generation of data,
model updating, optimum seeking



θ : designable parameter

x : observation from θ_h (simulation)

θ unknown

M : model

9
 M_h : imprecise

(locally reliable, at best)

Criterion for selecting E_{h+1} ?

(coverage of region, diagnostics, estimation)

$M_h \rightarrow M_{h+1}$?

from x_{h+1} , θ_h & diagnostics

local: regression model with
low order surface

global: doesn't fit,
rely on search.

10
Other statistical tools :

- for multivariate sparse data, projection pursuit,
- for highly correlated data (e.g. electrical measurements of 100 transistors on a chip) : factor analysis, ---
- for reliability prediction, use measured degradation, modeling & analysis.
-
-
-
-
-

DESIGNING FOR QUALITY
USING
COMPUTER EXPERIMENTATION

JEFF HOOPER

MAY 1988

OVERVIEW

**OBJECTIVE: DISCUSS THESE PAPERS IN
LIGHT OF SIGNIFICANT
DESIGN FOR QUALITY
EFFORTS USING PHYSICAL
EXPERIMENTATION**

- **LOSS FUNCTIONS AND ADJUSTMENT
PARAMETERS**
- **EXPERIMENTAL DESIGNS**
- **NOISE**
- **OPTIMIZATION**

INCREASING IMPORTANCE OF CADQ



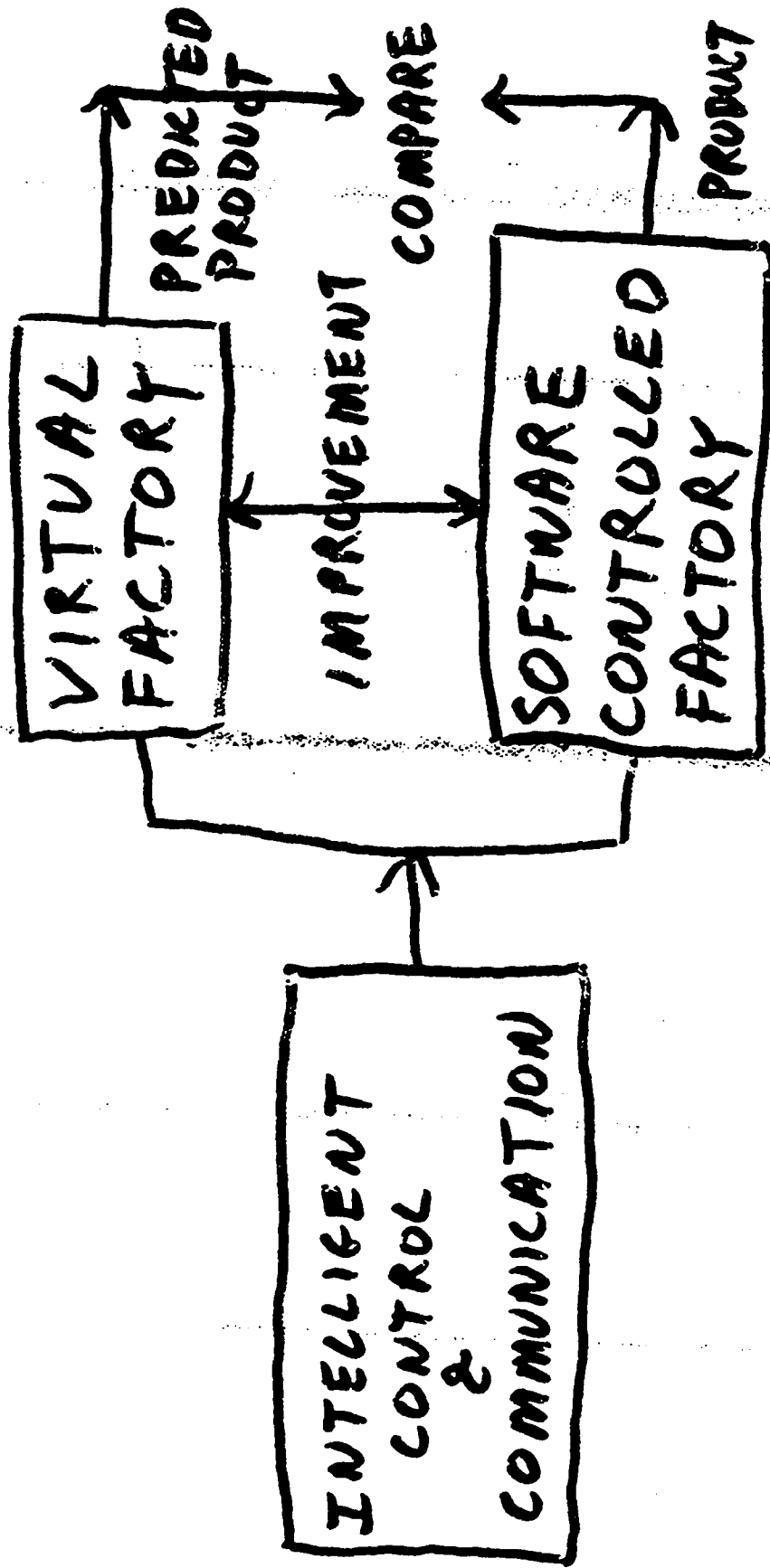
MORE PRODUCT AND PROCESS
DESIGN DONE ON COMPUTER

- AVAILABILITY OF COMPUTING POWER
- INCREASED PHYSICAL UNDERSTANDING
- LESS EXPENSIVE THAN PHYSICAL MODELS
- SHORTEN DESIGN INTERVAL
- EVALUATE MORE DESIGNS

DESIGN FOR QUALITY
INCREASING IN IMPORTANCE

- MOVE UPSTREAM TO PREVENT
QUALITY AND RELIABILITY PROBLEMS
- BUILD QUALITY IMPROVEMENT INTO
THE WAY WE DO THINGS

FUTURE IC FACTORY



DESIGN FOR QUALITY OBJECTIVE

**USE NOMINAL DESIGN VALUES TO
MINIMIZE THE INFLUENCE OF NOISE
ON THE PERFORMANCE OF THE DESIGN**

- **STATISTICAL DESIGN CENTERING
COMPUTER EXPERIMENTS**
- **ROBUST DESIGN
PHYSICAL EXPERIMENTS**

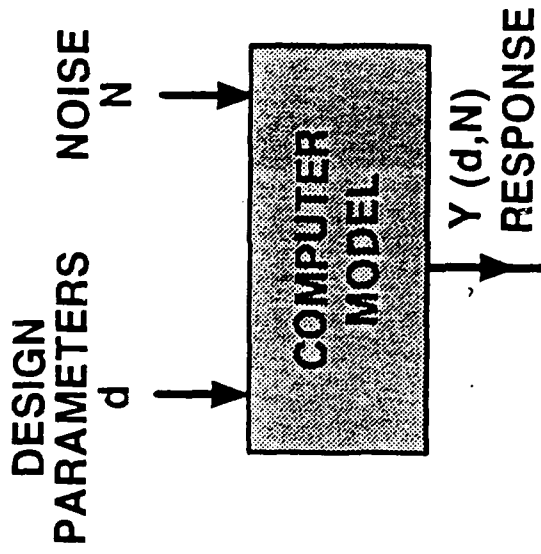
PHYSICAL VS. COMPUTER EXPERIMENTATION

	PHYSICAL	COMPUTER
• LOSS FUNCTION	SQUARED ERROR	0-1 (YIELD)
• ADJUSTMENT PARAMETERS	YES	NO
• NUMBER OF RUNS	SMALL	LARGE
• EXPERIMENTAL STAGES	FEW (1 OR 2)	MANY (SEQUENTIAL)
• EXPERIMENTAL DESIGN	SIMPLE	COMPLEX
• NOISE	SYSTEMATIC SAMPLING	RANDOM SAMPLING
• OPTIMIZATION	GLOBAL DESIGN OF EXPERIMENTS LINEAR MODELS	LOCAL OPTIMIZATION TECHNIQUES

LOSS FUNCTIONS AND ADJUSTMENT PARAMETERS

1. "SQUARED ERROR" VS 0-1 LOSS
 - ECONOMIC ADVANTAGES FOR BEING CLOSE TO TARGET
 - MAY REQUIRE MULTICRITERIA OPTIMIZATION
2. ADJUSTMENT PARAMETERS
 - HOW TO IDENTIFY
 - HOW TO CHOOSE A PERFORMANCE MEASURE

STATISTICAL DESIGN OPTIMIZATION



$$\begin{array}{ll} \text{MINIMIZE } E_N (W (Y(d,N))) \\ \text{s.t.} & d \in D \end{array}$$

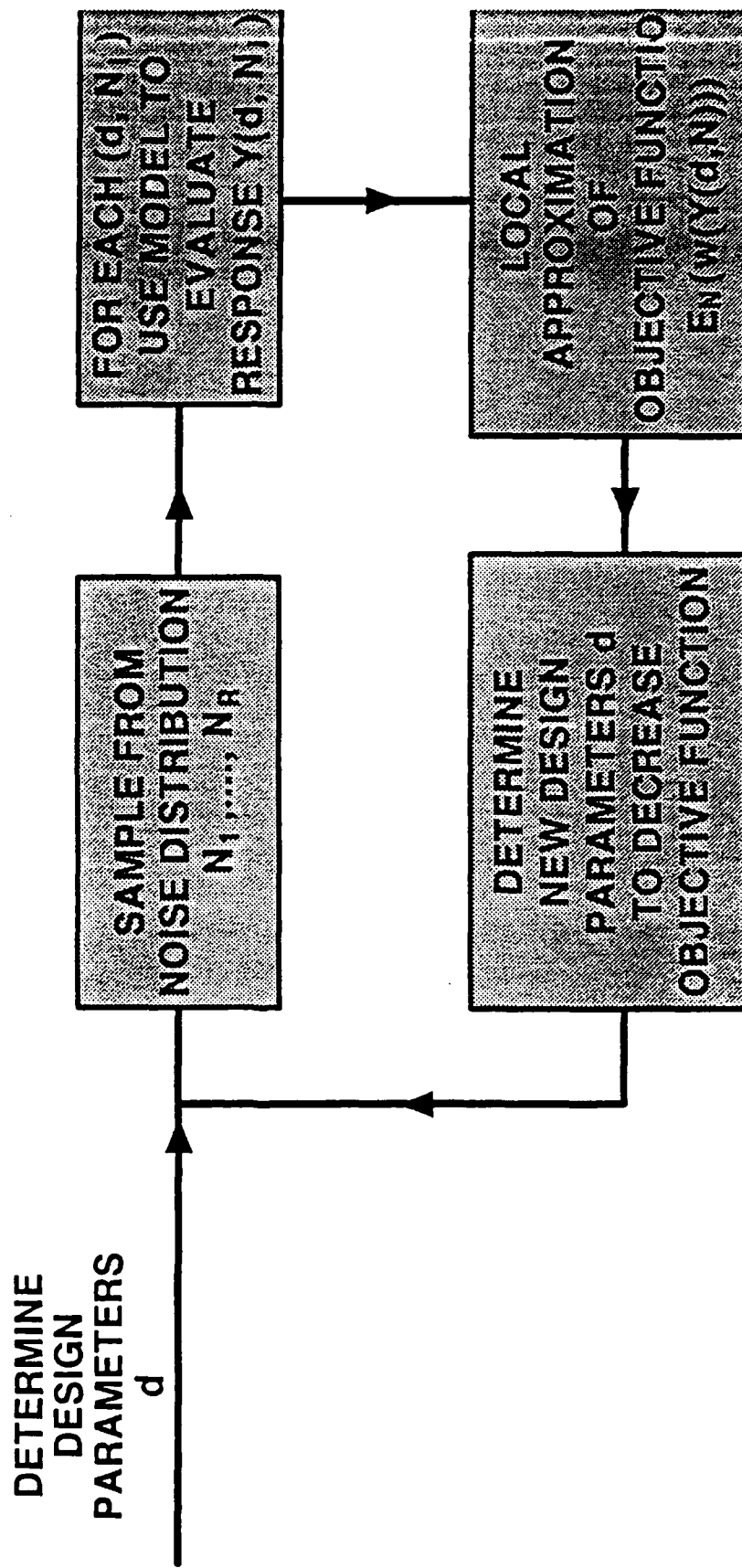
TWO BASIC APPROACHES

1. LOCAL APPROXIMATION OF THE OBJECTIVE FUNCTION
2. LOCAL APPROXIMATION OF COMPUTER MODEL



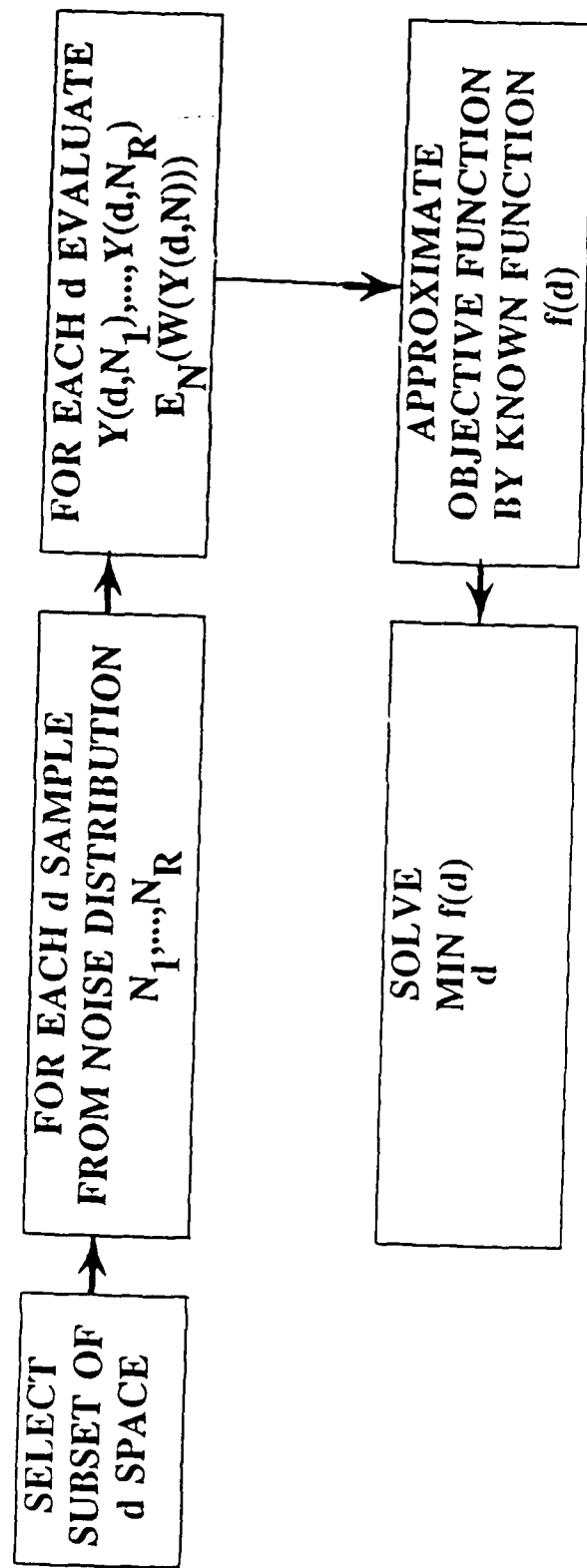
BASIC MONTE-CARLO METHOD

Local Approximation Of The Objective Function

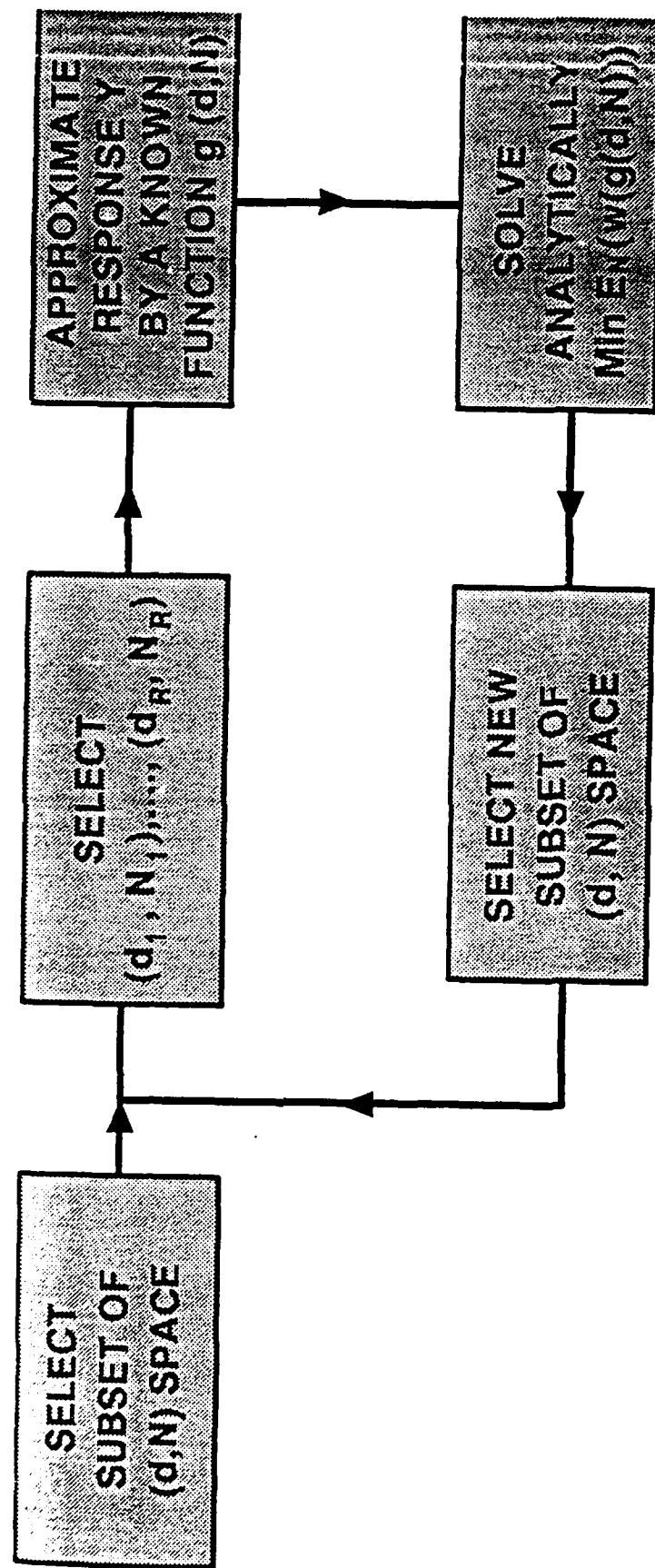


HGL74123.014

MODIFIED MONTE-CARLO METHOD
GLOBAL APPROXIMATION OF THE OBJECTIVE FUNCTION



LOCAL APPROXIMATION OF COMPUTER MODEL



HGL74123.013